CAUSAL INFERENCE WITH PARTIAL INTERFERENCE AND RIGHT CENSORED OUTCOMES

Sujatro Chakladar

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2018

Approved by:

Michael G. Hudgens

Chirayath M. Suchindran

Jianwen Cai

Michael Emch

Stephen R. Cole

ABSTRACT

Sujatro Chakladar: Causal Inference with Partial Interference and Right Censored Outcomes
(Under the direction of Michael G. Hudgens)

Interference arises when the outcome of one individual depends on the treatment status of another individual. Partial interference is a special case of interference where individuals can be partitioned into groups such that no interference occurs between groups but may occur within groups. In the absence of interference, inverse probability weighted (IPW) estimators are commonly used to draw inference about causal effect. Tchetgen Tchetgen and VanderWeele (2012) proposed a modified IPW estimator for different causal effects in the presence of partial interference. An extension of the Tchetgen Tchetgen and VanderWeele IPW estimator is proposed for the setting where the outcome is subject to right censoring using inverse probability of censoring weights (IPCW). Censoring weights are calculated using parametric frailty models. The large sample properties of the IPCW estimators are derived and simulation studies are presented demonstrating the estimators' performance in finite samples. The methods are illustrated using data from a cholera vaccination trial in Matlab, Bangladesh.

Unfortunately, IPW methods often suffer from a significant disadvantage due to the instability of propensity scores. The generalized computation algorithm formula (g formula) is a natural alternative for IPW estimators. Robins (1986) proposed the use of g-computation algorithm in the absence of interference to infer causal estimands of interest. Since then, the parametric g formula has been used for data with time varying confounding and exposure and also for time to event data (Robins 1987, Taubman et al. 2009, Westreich et al. 2012, Keil et al. 2014). An extension of the parametric g formula

is proposed when there is time to event data with right censoring and possible partial interference. Parametric frailty models are used to model the probability of an event. Derivation of large sample properties of the estimator is provided. Simulation studies show the operating characteristics of the method for finite samples. The cholera vaccination trial in Matlab, Bangladesh is used to illustrate the methods in a real scenario.

But both of these methods rely on the intrinsic assumption that the underlying models are correctly specified. If the treatment model is incorrect then the IPW/IPCW estimator will be inconsistent. Similarly, if the outcome model is incorrect then the parametric g formula will not be consistent. A doubly robust method is proposed to incorporate robustness under model misspecification so that the estimator is consistent even when only one of the two models is correctly specified. Large sample properties of the estimator are discussed. Finite sample performance of the method is also observed through simulation and the results are compared with the IPW/IPCW and parametric g formula. Finally, the doubly robust method is also applied to the cholera vaccine trial.

ACKNOWLEDGEMENTS

This work was supported by NIH grant R01 AI085073. I would like to start by thanking my dissertation committee members for their valuable input and support. In particular, suggestions by Dr. Chirayath M. Suchindran and Dr. Stephen R. Cole about time dependent censoring and parametric frailty models have been instrumental in my project. I am also grateful to Dr. Jianwen Cai, especially for her recommendations on the finite sample behavior of my estimates. Finally, I am also thankful to Dr. Michael Emch for his comments that have made me appreciative about the implication of my method on real data and in turn, have helped a great deal to interpret the results of my work. Apart from my committee members, I would also like to thank Dr. M. Elizabeth Halloran for her comments and suggestions on my first paper.

During the course of my research work, I have often found myself lost trying to tackle problems both pertaining to my research work and my personal life. I am eternally grateful to my adviser Dr. Michael G. Hudgens for always helping me through those tough times. He has taught me a great deal about how to handle complex problems, being methodical, and research work in general. I regard him as a great teacher and I consider myself very fortunate to be one of his students. His unwavering motivation and support have made my research possible. I sincerely thank him for his guidance.

I had the opportunity to be part of the Causal Inference with Interference (CIWI) research group which has been a great experience. Suggestions of my group members have often helped me make progress towards my goal. I would like to thank all CIWI members: Brian Barkley, Bradley Saul, Wen Wei Loh, Shaina Mitchell, Kayla Kilpatrick, and Bryan Blette for their contributions which have made research work a lot easier.

I would also like to thank my friends in North Carolina who have made my life here easier and friends outside of North Carolina who have enriched my world. In particular, to name a few, I am grateful towards Arkopal Choudhury, Arkaprava Roy, Aniket Bera, Suman Chakraborty, Sujayam Saha, Debraj Das, Pratyaydipta Rudra, Sayan Banerjee, Sayan Dasgupta, Abhishek Pal Majumder, Jyotishka Datta, Shalini Choudhury, Maxwell Datta, Subhamay Saha, Sapna Rao, Anwesha Goswami, Rinku Majumder, and Samarpan Majumder for always being there for me. I am thankful to Delilah for being in my life. Last but not the least, I am forever grateful to Sohini Raha for her unending support, constant enthusiasm, and for always guiding me through all the tough times that I encountered.

Finally, I would like to dedicate my thesis to my parents, Shyamal and Gouri Chakladar, my sister Soumita Das, and my nephew Shreyansh Das. I could not have managed my journey without their support.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER 1: LITERATURE REVIEW **1**

1.1 Causal Inference 1

1.1.1 Inverse Probability Weighting 3

1.1.2 G Formula 4

1.1.3 Doubly Robust Estimator 4

1.2 Interference 6

1.3 Censoring 10

1.4 Motivating Data 12

1.5 Summary 13

CHAPTER 2: CAUSAL INFERENCE WITH PARTIAL INTERFERENCE AND RIGHT CENSORED OUTCOMES USING IPCW ESTIMATORS **14**

2.1 Introduction 14

2.2 Methods 15

2.2.1 Estimands 15

2.2.2 Assumptions 18

2.2.3 Proposed Estimator 19

2.3 Simulation Study 23

2.4 Data Analysis 25

2.4.1 Cholera Vaccine Study and Analysis . . . 25
2.4.2 Results . . . 27
2.5 Discussion . . . 29
CHAPTER 3: PARAMETRIC G-FORMULA WITH PARTIAL INTERFERENCE AND RIGHT CENSORING . . . 35
3.1 Introduction . . . 35
3.2 Methods . . . 36
3.2.1 Estimands . . . 36
3.2.2 Assumptions . . . 39
3.2.3 Proposed Estimator . . . 40
3.3 Simulation . . . 46
3.4 Data . . . 48
3.4.1 Cholera Vaccine Study and Analysis . . . 48
3.4.2 Results . . . 50
3.5 Discussion . . . 52
CHAPTER 4: DOUBLY ROBUST ESTIMATION FOR DATA WITH PARTIAL INTERFERENCE AND RIGHT CENSORING . . . 58
4.1 Introduction . . . 58
4.2 Methods . . . 60
4.2.1 Estimands . . . 60
4.2.2 Assumptions . . . 63
4.2.3 IPCW and Parametric G Formula Estimators . . . 64
4.2.4 Proposed Estimator . . . 67
4.2.5 Properties of the Proposed Estimator . . . 68
4.3 Simulation . . . 74
4.4 Data Analysis . . . 78

4.4.1 Cholera Vaccine Study and Analysis 78
4.4.2 Results 81
4.5 Discussion 82
CHAPTER 5: CONCLUSION 91
APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 2 93
APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 3 95
APPENDIX C: TECHNICAL DETAILS FOR CHAPTER 4 97
BIBLIOGRAPHY 101

LIST OF TABLES

2.1 Results from simulation study described in Section 2.3. α denotes the allocation probability, $\mu(100, a, \alpha)$ is the true value of the target parameter for $a = 0, 1$; Bias is the average of $\mu(100, a, \alpha) - \hat{\mu}(100, a, \alpha)$ for $a = 0, 1$; ESE is the empirical standard error; ASE is the average of the sandwich variance estimates; and EC denotes the empirical coverage of the 95% Wald confidence intervals. 34

3.1 Results from simulation study described in Section 3.3. α denotes the allocation probabilities, $\mu(100, a, \alpha)$ is the true value of the survival probabilities at time point 100 for $a = 0, 1$; bias is the average of $\mu(100, a, \alpha) - m^{int}(100, a, \alpha, \hat{\boldsymbol{\omega}})$, ESE is the empirical standard error, ASE is the average of the sandwich variance estimators and EC denotes the empirical coverage of the 95% Wald type confidence intervals. 56

3.2 AIC (BIC) values for different baseline hazard functions corresponding to gamma, inverse Gaussian and Positive stable frailty distributions. 57

4.1 Results from simulation study described in Section 4.3. α denotes the allocation probabilities, $\mu(100, a, \alpha)$ is the true value of the target parameter for $a = 0, 1$; Bias is the average of $\mu(100, a, \alpha) - \hat{F}^{DR}(100, a, \alpha)$ for $a = 0, 1$; ESE is the empirical standard error, ASE is the average of the sandwich variance estimates and EC denotes the empirical coverage of the 95% Wald confidence intervals. 89

4.2 AIC (BIC) values for different baseline hazard functions corresponding to gamma, inverse Gaussian and positive stable frailty distributions. 90

LIST OF FIGURES

2.1 Absolute bias (left) and 95% confidence interval coverage (right) for different numbers of groups for $\alpha = 0.5$. The dotted line in the right plot corresponds to 95% coverage. 31

2.2 Estimated cumulative probability of cholera over time for vaccinated and unvaccinated for $\alpha = 0.3$ (left), $\alpha = 0.45$ (center) and $\alpha = 0.6$ (right) 32

2.3 Direct, indirect, total and overall effect estimates ($\times 1000$) for different allocation strategies at time $t = 1$ year. Indirect, total, and overall effects are with respect to $\alpha_2 = 0.4$. The shaded regions denote pointwise 95% confidence intervals of the estimates. 33

3.1 Estimated cumulative probability of cholera against time for vaccine and control for $\alpha = 0.3$ (left), $\alpha = 0.45$ (center) and $\alpha = 0.6$ (right) 54

3.2 Direct effect, indirect effect, total effect and overall effect estimates multiplied by 1000 for different allocation strategies at time $t = 1$ year. Indirect effects, total effects and overall effects are with respect to $\alpha_2 = 0.4$. The shaded region denotes the 95% confidence interval of the estimates. 55

4.1 Absolute biases of the doubly robust, parametric g and the IPCW estimators under different model misspecifications in control group. 84

4.2 Absolute biases of the doubly robust, parametric g and the IPCW estimators under different model misspecifications in treatment group. 85

4.3 Coverages of the doubly robust, parametric g and the IPCW estimators under different model misspecifications in control group. 86

4.4 Coverages of the doubly robust, parametric g and the IPCW estimators under different model misspecifications in treatment group. 87

4.5 Direct effect, indirect effect, total effect and overall effect estimates multiplied by 1000 for different allocation strategies at time $t = 1$ year. Indirect effects, total effects and overall effects are with respect to $\alpha_2 = 0.4$. The shaded region denotes the 95% confidence interval of the estimates. 88

CHAPTER 1: LITERATURE REVIEW

1.1 Causal Inference

Association does not always imply causation. Causal inference aims to address causality in statistical inference. Splawa-Neyman et al. (1923) put forward the idea of potential outcomes which has become a building block of causal inference. Potential outcomes or counterfactual outcomes are defined to be all possible outcomes for a study which are not necessarily observed and can be treated as missing data. For example, consider a randomized clinical trial with a vaccine and a placebo. Then, the outcome when an individual would be administered placebo is not observed if the individual was administered treatment. This gives rise to the concept of potential outcomes or counterfactuals. In terms of mathematics, denote by Y_i $(i = 1, \ldots, n)$, the binary outcome of n individuals. A_i represents the binary treatment status of individual i. A_i equals zero if the i-th individual receives placebo and equals one if said individual receives treatment. The potential outcome for individual i under treatment a is denoted by Y_i^a. So, if individual i receives placebo then $Y_i = Y_i^0$ and Y_i^1 is unobserved and if individual i receives treatment then $Y_i = Y_i^1$ and Y_i^0 is unobserved. This is termed to be causal consistency formally defined first in Gibbard and Harper (1976) and discussed in detail by Cole and Frangakis (2009). Applications of this representation can be found in various literature including (Haavelmo 1944, Robins and Greenland 1996, Pratt and Schlaifer 1988), epidemiology (Greenland and Robins 1986, Robins et al. 2000; 1992, Robins 2000) social and behavioral sciences (Sobel 1990; 1995, Willkinson 1999) and statistics (Rubin 2004, Pratt and Schlaifer 1984).

Rubin (1974, 1977, 1978) defined various causal effects in terms of counterfactuals.

For example, the average causal effect is defined as $E(Y^1) - E(Y^0)$. He extended the ideas for randomized experiments to non randomized studies. Most of the causal inference framework assumes the Stable unit treatment value assumption (SUTVA) put forth by Rubin (1980). The assumptions are-

1. There is no interference between individuals. That is, the treatment status of one individual in no way affects the outcome of another individual. Mathematically A_i has no effect on Y_j if $i \neq j$.

2. There is only one version of treatment and control.

Along with SUTVA, in the case of a conditional randomized experiment or in an observational setting, it is often assumed that given a set of measured covariates, the potential outcomes are independent of treatment. Denoting the set of measured confounders by $\mathbf{L}$, this assumption can be represented as $Y^a \perp\!\!\!\perp A|\mathbf{L}$. This assumption is often referred to as conditional exchangeability (Hernán and Robins (2006), Hernan and Robins (2010)). Conditional exchangeability fails when there exist unmeasured covariates affecting both the treatment and outcome. Hence this assumption is also known as no unmeasured confounding. Another assumption that is often found in causal literature is the assumption of positivity discussed by Westreich and Cole (2010). According to this assumption, $\Pr(A = a|\mathbf{L} = \mathbf{l}) > 0$ for $a \in (0, 1)$ when $\Pr(\mathbf{L} = \mathbf{l}) > 0$. i.e. if a particular value of the covariates has a positive probability of being observed in the data then given that value of the covariate, the probability of an individual being in the treatment as well as in he control group must be positive. These assumptions are discussed in depth by Hernán and Robins (2006). Under these assumptions, popular methods for estimating causal effects for observational studies include inverse probability treatment weighted (IPTW) estimator, parametric g formula and doubly robust estimator.

1.1.1 Inverse Probability Weighting

The idea of inverse probability weighting revolves around creating a pseudo population of individuals by weighting the study population with appropriate weights. The earliest known use of these IPW type estimators were suggested by Horvitz and Thompson (1952). These estimates have been used to estimate various causal effects like the average causal effect. For example, the IPTW estimate of $E(Y^a)$ is given to be

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a)Y_i}{\Pr(A_i=a|\mathbf{L_i})}$$

The weight for individual i is $\frac{I(A_i=a)}{\Pr(A_i=a|\mathbf{L_i})}$. The assumption of positivity ensures that the term in the denominator $\Pr(A_i=a|\mathbf{L_i})$ is greater than 0. Hence the estimator is well defined. Here, $\mathbf{L_i}$ is the vector of covariates for individual i. Under the assumptions discussed previously, it has been shown to be an unbiased estimator of the causal quantity of interest $E(Y^a)$ if $\Pr(A_i=a|\mathbf{L_i})$ is known. The quantity $\Pr(A_i=a|\mathbf{L_i})$ is termed as the propensity score (Rosenbaum and Rubin 1983). So, for individuals receiving treatment, in the estimation of $E(Y^0)$, they are assigned zero weight. Whereas in the estimation of $E(Y^1)$, they are assigned the weight of the inverse of the probability that they receive treatment given their individual covariates. However in an observational setting, the propensity scores are seldom known beforehand. Rosenbaum and Rubin (1983) suggested estimating the weights using a logit model. IPW estimators suffer from a serious drawback. When the propensity score is close to zero, then the estimator becomes large and computation becomes difficult (Little and Rubin 2014, Cole and Hernán 2008). Inclusion of too many variables in the model might give rise to these problems. Certain covariates might generate very low probability of treatment. This occurrence is sometimes called narrow strata in the literature (Lefebvre et al. 2008).

1.1.2 G Formula

The generalized computation algorithm formula (g formula) provides an alternative way to compute causal effects and is free from the aforementioned drawback of IPW estimators discussed in the previous section. Robins (1986) proposed the g formula to estimate causal effects. Using parametric outcome regression models along with the g formula gives rise to the parametric g formula. This method is a generalization of standardization (Hernán and Robins 2006). The assumptions discussed previously facilitate writing $E(Y^a)$ as being equal to $\int E(Y|A = a, \mathbf{L} = \mathbf{l})dF_{\mathbf{L}}(\mathbf{l})$. Again, the positivity assumption ensures that the parameter is well defined. So, an estimate of this causal quantity of interest $\hat{E}(Y^a)$ is obtained by empirically estimating the distribution of $\mathbf{L}$. The form of the estimator is given as follows-

$$\frac{1}{n}\sum_{i=1}^{n}\hat{E}(Y|A = a, \mathbf{L_i})$$

$\hat{E}(Y|A = a, \mathbf{L_i})$ is estimated using an outcome regression model. For example, in case of a binary outcome, a logit model might be appropriate for modeling Y conditional on A and $\mathbf{L}$. Parametric g formula has been proven to be particularly useful in adjusting for time varying confounders for time to event data (Young et al. 2011). Parameters of interest such as risk ratio (Taubman et al. 2009, Garcia-Aymerich et al. 2013, Cole et al. 2013) and hazard ratio (Westreich et al. 2012, Keil et al. 2014) have been calculated using this method. In all of these papers, pooled Logistic regression is the choice of parametric outcome model.

1.1.3 Doubly Robust Estimator

Both the IPW estimator and the parametric g formula operate under the obvious intrinsic condition that the underlying model is specified correctly. However, if the treatment model is specified incorrectly then the IPW will produce erroneous results (Lefebvre

et al. 2008, Cole and Hernán 2008). Similarly if the outcome model is specified incorrectly then the estimates obtained using the parametric g formula can not be trusted (Taubman et al. 2009). The doubly robust estimator incorporates robustness within the estimator in the sense that the doubly robust estimator will generate reasonable results even when only one of the treatment and outcome regression models are specified correctly. Doubly robust estimators have existed in literature for sample survey data (Cassel et al. 1977, Särndal et al. 2003). There have been a number of papers on doubly robust estimators applied for the missing data problem and causal inference (Kang and Schafer 2007, Bang and Robins 2005). Lunceford and Davidian (2004) put forth a doubly robust estimator for estimating causal effects. The estimator proposed was a weighted combination of the IPW estimator and the parametric g estimator. The estimator is as follows

$$\tilde{E}[Y^1] = n^{-1} \sum_{i=1}^{n} \frac{A_i Y_i - \{A_i - \hat{\Pr}(A_i = a|\mathbf{L_i})\}\hat{m}_1(\mathbf{L_i})}{\hat{\Pr}(A_i = a|\mathbf{L_i})}$$

and

$$\tilde{E}[Y^0] = n^{-1} \sum_{i=1}^{n} \frac{(1 - A_i) Y_i + \{A_i - \hat{\Pr}(A_i = a|\mathbf{L_i})\}\hat{m}_0(\mathbf{L_i})}{1 - \hat{\Pr}(A_i = a|\mathbf{L_i})}$$

where $\hat{m}_a(\mathbf{L_i}) = \hat{E}(Y|A = a, \mathbf{L_i})$ The authors showed large sample results for the doubly robust estimator. They showed that the estimate is consistent asymptotically normal and they calculated large sample variance as well. The authors also showed the doubly robust property of the estimate. Research is still ongoing for doubly robust estimators. Cao et al. (2009) improved upon the doubly robust estimator. When both the models are misspecified, doubly robust estimators give rise to significant biases. They also give rise to biases for the case when the estimated propensity score becomes close to 0. The authors tried to address this issue by modifying the existing doubly robust estimator and showed that they perform better in the said situation. Funk et al. (2011) provides a good theoretical and practical explanation of the properties of the doubly robust estimator for causal outcomes of interest.

1.2 Interference

Interference is said to be present in a data when the the treatment status of one individual affects the outcome of another individual (Cox 1958). An example of this might be data on infectious diseases. In this setting, whether a person gets infected or not might be affected by the treatment status of another individual (Halloran and Struchiner 1991). Partial interference is a special case of interference. Partial interference occurs when we have a partition of the data such that interference is observed within the individuals of a group but not between the individuals in different groups (Sobel 2006). Data might show traits of partial interference if there is a clear demographic separation between groups of individuals based on geography, society, or temporality. Therefore, interference can introduce possible indirect effects of interest which are termed spillover effects, or peer effects. These effects have been discussed in various fields including criminology (Sampson 2010, Verbitsky-Savitz and Raudenbush 2012), developmental psychology (Duncan et al. 2005, Foster 2010), econometrics (Sobel 2006, Manski 2013), education (Hong and Raudenbush 2006, Vanderweele et al. 2013), imaging (Luo et al. 2012), political science (Sinclair et al. 2012, Bowers et al. 2013), social media and network analysis (VanderWeele and An 2013, Toulis and Kao 2013), (Eckles et al. 2014, Kramer et al. 2014), sociology (Aronow and Samii 2017), and spatial analyses (Zigler et al. 2012, Graham et al. 2013).

Various literature propose methods for calculating interference effects in a randomized setting (Rosenbaum 2007, Hudgens and Halloran 2008, Baird et al. 2018, Eckles et al. 2016). Hudgens and Halloran (2008) presented estimands of direct, indirect, total and overall effects and proved the relations between them which was first discussed by Halloran and Struchiner (1991). Modifying previous notation, suppose data now has m groups of individuals, with n_i individuals per group for $i = 1, \ldots, m$. Denote by A_{ij}, the indicator of treatment status of individual j in group i i.e., if individual j in group i receives treatment then $A_{ij} = 1$ and otherwise $A_{ij} = 0$. Also, let the vector of treatment indicator of group i be denoted by $\mathbf{A}_i$ and that of group i except for individual j be denoted by

$\mathbf{A}_{i,-j}$. So, $\mathbf{A}_i = (A_{i1}, A_{i2}, ..., A_{in_i})$ and $\mathbf{A}_{i,-j} = (A_{i1}, A_{i2}, ..., A_{ij-1}, A_{ij+1}..., A_{in_i})$. Assume that possible realizations of $\mathbf{A}_i$ and $\mathbf{A}_{i,-j}$ are denoted by $\mathbf{a_i}$ and $\mathbf{a_{i,-j}}$ respectively. The outcome of individual j in group i is denoted by Y_{ij} $(i = 1, \ldots, m,\ j = 1, \ldots, n_i)$. The potential outcome for individual j in group i under treatment a for the individual and treatment vector $\mathbf{a}_{i,-j}$ for the rest of the individual in group i is denoted by $Y_{ij}(a, \mathbf{a}_{i,-j})$. Since there are only two versions of treatments, if there are n individuals in a group, then the set of all possible group treatment assignments consist of 2^n elements and that set is denoted by $\mathcal{A}(n)$ for $n = 1, 2, \ldots$. The vector of covariates for subject j in group i is denoted by $\mathbf{L}_{ij}$ and the matrix of covariates for all subjects in group i is denote by $\mathbf{L}_i$, i.e. $\mathbf{L}_i = (\mathbf{L}_{i1}, \mathbf{L}_{i2}, \cdots, \mathbf{L}_{in_i})$. Hudgens and Halloran (2008) defined individual average potential outcome to be

$$\bar{Y}_{ij}(a, \alpha) = \sum_{\mathbf{a}_{i,-j} \in \mathcal{A}(n_i - 1)} Y_{ij}(a, \mathbf{a}_{i,-j}) \pi(\mathbf{a}_{i,-j}, \alpha)$$

and the marginal individual average potential outcome to be

$$\bar{Y}_{ij}(\alpha) = \sum_{\mathbf{a}_i \in \mathcal{A}(n_i)} Y_{ij}(\mathbf{a}_i) \pi(\mathbf{a}_i, \alpha)$$

where α denotes the group allocation strategies (Hong and Raudenbush 2006, Sobel 2006, Tchetgen and VanderWeele 2012, Hudgens and Halloran 2008). Specifically, according to the "Bernoulli" treatment allocation strategy discussed in Tchetgen Tchetgen and Vanderweele (2012) α can be interpreted as the probability with which an individual receives treatment independently of others. Also, $\pi(\mathbf{a}_{i,-j}, \alpha)$ denote the conditional probability that the treatment assignment for the i^{th} group except for the j^{th} individual is $\mathbf{a}_{i,-j}$ given that the j^{th} individual in the i^{th} group receives treatment a under allocation strategy α. In terms of probability, $\pi(\mathbf{a}_{i,-j}, \alpha) = \Pr(\mathbf{A}_{i,-j} = \mathbf{a}_{i,-j} | A_{i,j} = a)$. Then,

$\pi(\mathbf{a}_{i,-j}, \alpha) = \Pi_{k=1,k\neq j}^{n_i} \alpha^{a_{ik}}(1-\alpha)^{1-a_{ik}}$. Similarly, let $\pi(\mathbf{a}_i, \alpha)$ denote the conditional probability that the treatment assignment for the i^{th} group is $\mathbf{a}_i$ under allocation strategy α. In terms of probability, $\pi(\mathbf{a}_i, \alpha) = \Pr(\mathbf{A}_i = \mathbf{a}_i)$. Then, $\pi(\mathbf{a}_i, \alpha) = \Pi_{k=1}^{n_i} \alpha^{a_{ik}}(1-\alpha)^{1-a_{ik}}$. The population average potential outcome is defined as

$$\mu(a, \alpha) = E\left\{ n_i^{-1} \sum_{j=1}^{n_i} \bar{Y}_{ij}(a, \alpha) \right\}$$

and the marginal population average potential outcome is defined as

$$\mu(\alpha) = E\left\{ n_i^{-1} \sum_{j=1}^{n_i} \bar{Y}_{ij}(a, \alpha) \right\}$$

Then, according to Halloran and Struchiner (1995) and Hudgens and Halloran (2008) the population average direct causal effect is a measure of the direct difference in effects between vaccinated and unvaccinated individuals. It is given by $\mu(0, \alpha) - \mu(1, \alpha)$. The population average indirect casual effect is a measure of the herd spillover effect and is the difference between the outcome of two unvaccinated individual under two different allocation strategies. It is given by $\mu(0, \alpha_1) - \mu(0, \alpha_2)$ for allocation strategies α_1 and α_2. Population average total effect is a combination of both the direct and the indirect effects. It is obtained by taking the difference between the population average potential outcome of untreated individual at allocation level α_1 and treated individuals at allocation level α_2, i.e. $\mu(0, \alpha_1) - \mu(1, \alpha_2)$. Finally population average overall effect is obtained by taking the difference between the average potential outcomes of individuals at allocation level α_1 and individuals at allocation level α_2, i.e. $\mu(\alpha_1) - \mu(\alpha_2)$. If there is no interference present, then the direct effect would be equal to the total effect and the indirect effect would be 0. A two stage stratified interference method is applied and unbiased estimates of the parameters and also the variance estimator given the first stage of randomization is obtained given various assumptions.

It is not possible to perform randomized experiments in all cases. For example, there

might be ethical issues in implementing treatments selectively to randomized individuals. Tchetgen Tchetgen and VanderWeele (TV) (2012) suggested using inverse probability weighted (IPW) estimators for causal effects for observational data in such cases when partial interference might be present within the data. The estimators were constructed using group propensity scores instead of individual propensity scores. Perez-Heydrich et al. (2014) and Liu et al. (2016) showed large sample properties of these estimates. The estimate for the group level average potential outcome and marginal group level average potential outcome are given respectively by

$$\hat{Y}_i^{TV}(a,\alpha) = n_i^{-1}\sum_{j=1}^{n}\frac{\pi(\mathbf{A}_{i,-j};\alpha)I(A_{ij}=a)Y_{ij}}{\Pr(\mathbf{A}_i|\mathbf{L}_i)}$$

and

$$\hat{Y}_i^{TV}(t,\alpha) = n_i^{-1}\sum_{j=1}^{n}\frac{\pi(\mathbf{A}_i;\alpha)Y_{ij}}{\Pr(\mathbf{A}_i|\mathbf{L}_i)}$$

Using the estimators proposed by Tchetgen and VanderWeele (2012), Perez-Heydrich et al. (2014) went on to calculate group propensity scores via modeling the probability of participation using a mixed effects model. Using estimating equation representation and results from Stefanski and Boos (2002), asymptotic variance estimators were calculated and their estimators were given.

Liu et al. (2018) extends doubly robust estimators to the case with partial interference. The form of the group average potential outcome estimate is as follows-

$$\hat{Y}_i^{DR}(a,\alpha) = n_i^{-1}\sum_{j=1}^{n_i}[\frac{I(A_{ij}=a)\{Y_{ij}(\mathbf{A}_i) - m_{ij}(\mathbf{A}_i, L_{ij}, \hat{\beta})\}\pi(\mathbf{A}_{i,-j};\alpha)}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\hat{\omega})}$$

$$+\sum_{\mathbf{a}_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},L_{ij},\hat{\beta})\pi(\mathbf{a}_{i,-j};\alpha)]$$

Large sample properties of the estimator are also discussed.

1.3 Censoring

In settings where the outcome of interest is survival time (e.g., time until infection), the outcome is typically subject to (right) censoring due to study completion or participant drop-out. Various semi parametric and parametric methods has been explored in order to analyze clustered survival data. Glidden and Vittinghoff (2004) provides a comparison of various methods of formulating the hazard function. Holt and Prentice (1974) put forth a stratified Cox model with hazard function $g_{ij}(.)$ having the following form

$$g_{ij}(t|\mathbf{L}_{ij}) = g_{0i}(t) \exp(\mathbf{L}_{ij}^T \boldsymbol{\gamma}),$$

where the group specific baseline hazards $g_{0i}(.)$ are completely unspecified. Multiple failure times for a single subject, which can be interpreted as correlated survival data, has also been analyzed using this method and the asymptotic properties of the estimator have been provided (Wei et al. 1989). However, this did not facilitate between group comparisons and Holt and Prentice (1974) suggested separate parameters to be estimated for the different groups, i.e.,

$$g_{ij}(t|\mathbf{L}_{ij}) = \xi_i g_0(t) \exp(\mathbf{L}_{ij}^T \boldsymbol{\gamma}),$$

Specifically, the authors considered the model corresponding to $g_0(t) = 1$ (exponential) and $g_0(t) = t^{\eta-1}$. Pankratz et al. (2005) considered adding a random effect or frailty term e_i keeping the baseline hazard unspecified as follows

$$g_{ij}(t|\mathbf{L}_{ij}) = g_0(t) \exp(\mathbf{L}_{ij}^T \boldsymbol{\gamma} + e_i).$$

Their method made use of Laplace approximation for calculating the maximum likelihood estimator for the general random effect proportional hazards model. Vaupel et al. (1979) first used the term frailty in the context of mortality studies. Other methods used in

the survival literature for analyzing correlated data use Bayesian techniques (Clayton 1991) like Gibbs sampling (Gauderman and Thomas 1994, Korsgaard et al. 1998), and Monte-Carlo EM algorithm (Li and Thompson 1997); and frequentist approaches (Li and Zhong 2002) like penalized likelihood maximization (Therneau et al. 2003).

Parametric frailty models have also been used to draw inference about right censored correlated data (Lancaster 1979, Hougaard 1984). Gutierrez et al. (2002) gives a detailed description of the various use of parametric frailty model in literature for analyzing right censored data. The general form of the parametric frailty model is given by

$$g_{ij}(t|\mathbf{L}_{ij}, e_i) = g_0(\lambda, t)e_i \exp\left(\mathbf{L}_{ij}^T\boldsymbol{\gamma}\right)$$

where g_0 is the baseline hazard function, λ is the parameter of the baseline hazard function, e_i is a random component following density $f_e(e_i;\theta_r)$, t is the time to event and $\boldsymbol{\gamma}$ is the vector of coefficients (Munda et al. 2012). The baseline hazard function is assumed to have a parametric form. Various distributions like exponential, Weibull etc. are used as the distribution of baseline hazard. The random component is known as the frailty term.

Marginal structural models have been used for estimation of causal effects for right censored data specifically for time dependent confounders (Robins et al. 2000). The parameters are typically obtained by IPW. For example, Hernán et al. (2000) used marginal structural Cox proportional hazards model to estimate the effect of a time varying exposure named zidovudine on the survival of HIV-positive men. Robins and Finkelstein (2000) used inverse probability censoring weighted (IPCW) version of log rank test statistic to show that bactrim therapy improved survival of AIDS patients. The IPCW method has been rather popular for causal inference in right censored data (Cole and Hernán 2008, Cain and Cole 2009, Howe et al. 2011).

1.4 Motivating Data

A good motivating example for data with interference and right censoring is the cholera vaccine study in Matlab from a cholera vaccine trial in Matlab, Bangladesh (Ali et al. 2005). The data from this study were from children between the ages of 2 and 15 and women. The range of years through which data were collected was from 1985 to 1988. There were three available treatment groups and all the participants were randomized to one of these three groups. These treatment groups are B subunit-killed whole-cell oral cholera vaccine, killed whole-cell-only cholera vaccine and E. coli K12 placebo. The two vaccines were considered the same here for analysis purposes. The inclusion of a participant was subject to the condition that the individual received two or more doses of the vaccine. Non-participants were tracked using a vector of participation. The participation vector comprised of indicators of participation fro each individual. The vaccine and placebo were given from January 1985 to May 1985. Three centers were set up in the Matlab area for administering the dosages and all centers were used as surveillance centers as well. A total of 121,982 individuals were included. Perez-Heydrich et al. (2014) have shown in their paper that interference is present. Related individuals lived in clustered sets of houses called baris. There were a total of $6,415$ baris. Demographic separation was used to categorize all the baris in the study to different groups (neighborhoods). The total number of these neighborhoods were prespecified to be 700. The assumption of partial interference treated these neighborhoods as groups such that cholera can be transmitted from one individual to another within a neighborhood but not between individuals of two separate neighborhoods. Studies discovered that the probability of cholera among unvaccinated individuals was less for the neighborhoods with higher vaccination coverage (Ali et al. 2005, Emch et al. 2006).

All of these studies have failed to adjust for censoring. The time of cholera is observed only for a few individuals in the data and the rest are unobserved. The definition of censoring, in this case, is if an individual did not observe an event within the study

period, migrated elsewhere from the study location or died during the follow up period.

1.5 Summary

Methods for estimating various causal effects have been in the literature for randomized as well as observational studies under SUTVA. Some of these methods have been extended for the case where the data show traits of partial interference. Methods have been proposed for calculating various causal effects of interest when the data exhibit right censoring. Three popular methods include the IPW method, parametric g formula, and the doubly robust estimation method. This research aims at extending these ideas towards the case when there is partial interference as well as right censoring. Section 2, Section 3, and Section 4 explore corresponding methods using IPW, parametric g formula, and doubly robust estimation, respectively. The methods are discussed and theoretical results are given to show that the estimators are consistent and asymptotically normal. Simulation studies are performed for all the three methods to demonstrate the efficacy of the methods. All of these methods are applied to the Matlab cholera vaccination study and the results are compared to previous research.

CHAPTER 2: CAUSAL INFERENCE WITH PARTIAL INTERFERENCE AND RIGHT CENSORED OUTCOMES USING IPCW ESTIMATORS

2.1 Introduction

Interference arises when the outcome of one individual depends on the treatment status of another individual (Cox 1958). For example, in the setting of infectious diseases, whether one individual receives a vaccine may affect whether another individual becomes infected (Halloran and Struchiner 1991). Partial interference is a special case of interference where individuals can be partitioned into groups such that interference does not occur between individuals in different groups but may occur between individuals in the same group (Sobel 2006). Partial interference might be a reasonable assumption if groups of individuals are sufficiently separated geographically, socially, and/or temporally. Effects due to interference, also known as spillover effects or peer effects, are of interest in many areas, including criminology (Sampson 2010, Verbitsky-Savitz and Raudenbush 2012), developmental psychology (Duncan et al. 2005, Foster 2010), econometrics (Sobel 2006, Manski 2013), education (Hong and Raudenbush 2006, Vanderweele et al. 2013), imaging (Luo et al. 2012), political science (Sinclair et al. 2012, Bowers et al. 2013), social media and network analysis (VanderWeele and An 2013, Toulis and Kao 2013, Kramer et al. 2014, Eckles et al. 2014), sociology (Aronow and Samii 2017), and spatial analyses (Zigler et al. 2012, Graham et al. 2013).

Inferential methods about spillover effects have been developed for randomized experiments (Rosenbaum 2007, Hudgens and Halloran 2008, Eckles et al. 2016, Baird et al. 2018). However, in some settings it may not be feasible or ethical to randomize groups

or individuals to different treatment or exposure conditions. In the observational setting, Tchetgen Tchetgen and VanderWeele (henceforth TV) (2012) proposed inverse probability weighted (IPW) estimators for different types of causal effects when there may be partial interference. Large sample properties of these IPW estimators were considered by Perez-Heydrich et al. (2014) and Liu et al. (2016).

In settings where the outcome of interest is a time to event, the outcome may be subject to right censoring due to study completion or participant drop-out. In the absence of interference, censoring is often accommodated by using inverse probability of censoring weights along with inverse probability treatment weights (Robins et al. 2000, Hernán et al. 2000, Robins and Finkelstein 2000, Cole and Hernán 2008, Cain and Cole 2009, Howe et al. 2011). In this section, an extension of the TV IPW estimators is considered for the setting where there may be partial interference and the outcome is subject to right censoring using inverse probability of censoring weights (IPCW).

The outline of this section is as follows. The proposed methods are developed in Section 2.2. In Section 2.3 simulation results are presented demonstrating the empirical performance of the proposed methods in finite sample settings. In Section 2.4 the methods are used to analyze a cholera vaccine study of over 90,000 individuals in Matlab, Bangladesh. Section 2.5 concludes with a discussion.

2.2 Methods

2.2.1 Estimands

Suppose data are observed from m groups of individuals, with n_i individuals per group for $i = 1, \ldots, m$. Let $A_{ij} = 1$ if individual j in group i receives treatment and $A_{ij} = 0$ otherwise. Let $\mathbf{A}_i = (A_{i1}, A_{i2}, ..., A_{in_i})$ and $\mathbf{A}_{i,-j} = (A_{i1}, A_{i2}, ..., A_{ij-1}, A_{ij+1}..., A_{in_i})$. Let $\mathbf{a}_i$ and $\mathbf{a}_{i,-j}$ denote possible realizations of $\mathbf{A}_i$ and $\mathbf{A}_{i,-j}$, and let $\mathcal{A}(n)$ denote the set of all possible 2^n treatments for a group size of $n = 1, 2, \ldots$. Assume partial interference and denote the potential time to event for individual j in group i if, possibly counter to

fact, group i receives treatment $\mathbf{a}_i$ by $T_{ij}(\mathbf{a}_i)$. The notation $T_{ij}(\mathbf{a}_i)$ reflects the partial interference assumption, i.e., the potential outcome of individual j in group i does not depend on the treatment of individuals outside group i. Below the notation $T_{ij}(a, \mathbf{a}_{i,-j})$ is sometimes used to make explicit the treatment for individual j and the treatment for all other individuals in group i. Let $\mathbf{T}_i(.) = \{T_{ij}(\mathbf{a}_i) : \mathbf{a}_i \in \mathcal{A}(n_i), j = 1, 2, \cdots, n_i\}$ denote the set of all potential event times for individuals in group i. Suppose the event times are subject to right censoring, e.g., due to loss to follow-up or study completion. Let C_{ij} denote the potential censoring times for individual j in group i. Assume that treatment has no effect on the censoring times, i.e., C_{ij} does not depend on $\mathbf{a}_i$. Let $\Delta_{ij} = 1$ if $T_{ij}(\mathbf{A}_i) \leq C_{ij}$ and $\Delta_{ij} = 0$ otherwise, and let $X_{ij} = \min(T_{ij}(\mathbf{A}_i), C_{ij})$. Define $\mathbf{X}_i = (X_{i1}, X_{i2}, \cdots, X_{in_i})$ and $\boldsymbol{\Delta}_i = (\Delta_{i1}, \Delta_{i2}, \cdots, \Delta_{in_i})$. Denote by $\mathbf{L}_{ij}$ the vector of baseline covariates for subject j in group i and by $\mathbf{L}_i$, the baseline matrix of covariates for all subjects in group i, i.e., $\mathbf{L}_i = (\mathbf{L}_{i1}, \mathbf{L}_{i2}, \cdots, \mathbf{L}_{in_i})$. The group sizes n_i are assumed to be random variables included in the baseline covariates $\mathbf{L}_{ij}$. Assume that the m groups are randomly sampled from an infinite superpopulation of groups such that the observed data are m i.i.d. copies of $\mathbf{O}_i = (\mathbf{L}_i, \mathbf{A}_i, \mathbf{X}_i, \boldsymbol{\Delta}_i)$.

In the absence of interference, treatment effects are typically defined as contrasts in mean potential outcome for different counterfactual scenarios, e.g., the average treatment effect is usually defined as the difference in the mean potential outcomes had all individuals received treatment versus had no individuals received treatment. Similarly, in the setting where there is partial interference, causal effects may be defined as contrasts in mean potential outcomes for different counterfactual scenarios (Hong and Raudenbush 2006, Sobel 2006, Hudgens and Halloran 2008, Tchetgen and VanderWeele 2012). Here we consider counterfactual scenarios where the marginal probability that an individual receives treatment, $\Pr_\alpha(A_{ij} = 1)$, equals α for different values of $\alpha \in (0, 1)$. The notation $\Pr_\alpha(\cdot)$ indicates that the probability corresponds to the distribution under the counterfactual scenario. Specifically, we consider the Bernoulli treatment allocation strategy (or policy)

described in TV is considered wherein individuals independently select treatment with probability α. Let $\pi(\mathbf{a}_i, \alpha)$ denote the probability that group i receives treatment $\mathbf{a}_i$ under Bernoulli allocation strategy α. That is, $\pi(\mathbf{a}_i, \alpha) = \Pr_\alpha(\mathbf{A}_i = \mathbf{a}_i) = \prod_{k=1}^{n_i} \alpha^{a_{ik}}(1-\alpha)^{1-a_{ik}}$. Similarly let $\pi(\mathbf{a}_{i,-j}, \alpha) = \Pr_\alpha(\mathbf{A}_{i,-j} = \mathbf{a}_{i,-j} | A_{ij} = a) = \prod_{k=1,k\neq j}^{n_i} \alpha^{a_{ik}}(1-\alpha)^{1-a_{ik}}$.

The causal estimands of interest defined below are contrasts in the risk of having an event by time t for different combinations of treatment a and allocation strategies α. To define these estimands, let

$$\bar{F}_{ij}(t, a, \alpha) = \sum_{\mathbf{a}_{i,-j} \in \mathcal{A}(n_i - 1)} I\{T_{ij}(a, \mathbf{a}_{i,-j}) \leq t\} \pi(\mathbf{a}_{i,-j}, \alpha),$$

and

$$\bar{F}_{ij}(t, \alpha) = \sum_{\mathbf{a}_i \in \mathcal{A}(n_i)} I\{T_{ij}(\mathbf{a}_i) \leq t\} \pi(\mathbf{a}_i, \alpha).$$

In words, $\bar{F}_{ij}(t, a, \alpha)$ is the probability that individual j in group i will have an event by time t when receiving treatment a and the group adopts policy α. Likewise, $\bar{F}_{ij}(t, \alpha)$ is the probability that individual j in group i will have an event by time t when the group adopts allocation strategy α. Denote the group average risks by $\bar{F}_i(t, a, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \bar{F}_{ij}(t, a, \alpha)$ and $\bar{F}_i(t, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \bar{F}_{ij}(t, \alpha)$. Let $\mu(t, a, \alpha) = E_\alpha\{\bar{F}_i(t, a, \alpha)\}$ and $\mu(t, \alpha) = E_\alpha\{\bar{F}_i(t, \alpha)\}$ where $E_\alpha\{.\}$ denotes the expected value under the counterfactual setting when policy α is adopted in the super population of groups. In the cholera vaccine study described in Section 2.4, $\mu(t, a, \alpha)$ denotes the average risk of acquiring cholera by time t when an individual receives treatment a and other individuals receive vaccine with probability α. Various effects of treatment can be defined by contrasts in $\mu(t, a, \alpha)$ and $\mu(t, \alpha)$. The direct effect is obtained by comparing the probability of an event when an individual receives treatment versus when not receiving treatment for a fixed allocation strategy. In particular, the direct effect at time t corresponding to policy α is defined to be $DE(t, \alpha) = \mu(t, 0, \alpha) - \mu(t, 1, \alpha)$. The indirect (or spillover) effect is the difference in the probability of an event by time t for two different policies when

the individual does not receive treatment. Specifically, the indirect effect is given by $IE(t, \alpha_1, \alpha_2) = \mu(t, 0, \alpha_1) - \mu(t, 0, \alpha_2)$ for allocation strategies α_1 and α_2. An indirect effect can analogously be defined when an individual is vaccinated. The total effect is defined as the difference between the probability of an event by time t when an individual does not receive treatment under policy α_1 and when an individual receives treatment under policy α_2, i.e., $TE(t, \alpha_1, \alpha_2) = \mu(t, 0, \alpha_1) - \mu(t, 1, \alpha_2)$. Finally, the overall effect is the difference between the probability of an event by time t for policy α_1 versus α_2, i.e., $OE(t, \alpha_1, \alpha_2) = \mu(t, \alpha_1) - \mu(t, \alpha_2)$.

2.2.2 Assumptions

Assume the following for all $\mathbf{a}_i \in \mathcal{A}(n_i)$,

I) Conditional independence: $\mathbf{A}_i \perp\!\!\!\perp \mathbf{T}_i(.)|\mathbf{L}_i$,

II) Positivity: $\Pr(\mathbf{A}_i = \mathbf{a}_i|\mathbf{L}_i = \mathbf{l}) > 0$ for all $\mathbf{a}_i \in \mathcal{A}(n_i)$ and $\mathbf{l}$ such that $\Pr(\mathbf{L}_i = \mathbf{l}) > 0$,

III) Conditional independent censoring: $\mathbf{C}_i \perp\!\!\!\perp \{\mathbf{T}_i(.), \mathbf{A}_i\}|\mathbf{L}_i$.

Assumption I states that the potential event times for individuals within the same group are conditionally independent of the actual treatment received by the group given covariates; this is a group-level generalization of the usual individual-level no unmeasured confounders assumption often made in the absence of interference. Positivity assumes that each group has a non-zero probability of being assigned every possible treatment combination given covariates for the group. Assumption III states that the potential event times for individuals in the same group and the observed group treatment are conditionally independent of the censoring times given covariates. It is straightforward to adapt the methods below to allow for a weaker version of assumption III whereby the censoring times are conditionally independent of the potential outcomes given group treatment and covariates.

2.2.3 Proposed Estimator

In the absence of censoring, the IPW estimator proposed by TV can be used to draw inference about $\mu(t, a, \alpha)$ and $\mu(t, \alpha)$. In particular, letting $Y_{ij} = I(X_{ij} \leq t)$, the TV IPW estimators are $\hat{\mu}^{TV}(t, a, \alpha) = m^{-1} \sum_{i=1}^{m} \hat{F}_i^{TV}(t, a, \alpha)$ and $\hat{\mu}^{TV}(t, \alpha) = m^{-1} \sum_{i=1}^{m} \hat{F}_i^{TV}(t, \alpha)$ where

$$\hat{F}_i^{TV}(t, a, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \frac{\pi(\mathbf{A}_{i,-j}; \alpha) I(A_{ij} = a) Y_{ij}}{\Pr(\mathbf{A}_i | \mathbf{L}_i, \hat{\boldsymbol{\beta}})}, \quad \hat{F}_i^{TV}(t, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \frac{\pi(\mathbf{A}_i; \alpha) Y_{ij}}{\Pr(\mathbf{A}_i | \mathbf{L}_i, \hat{\boldsymbol{\beta}})},$$

and $\hat{\boldsymbol{\beta}}$ is an estimator of the vector of parameters for the propensity model $\Pr(\mathbf{A}_i | \mathbf{L}_i, \boldsymbol{\beta})$. Details of the propensity model are discussed in the next sections.

In the presence of censoring, the following extension of the TV IPW estimators is proposed: $\hat{\mu}(t, a, \alpha) = m^{-1} \sum_{i=1}^{m} \hat{F}_i(t, a, \alpha)$ and $\hat{\mu}(t, \alpha) = m^{-1} \sum_{i=1}^{m} \hat{F}_i(t, \alpha)$ where

$$\hat{F}_i(t, a, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \frac{\pi(\mathbf{A}_{i,-j}; \alpha) I(A_{ij} = a) I(\Delta_{ij} = 1) I(X_{ij} \leq t)}{\Pr(\mathbf{A}_i | \mathbf{L}_i, \hat{\boldsymbol{\beta}}) \Pr(\Delta_{ij} = 1 | \mathbf{L}_i, X_{ij}, \hat{\boldsymbol{\gamma}})},$$

$$\hat{F}_i(t, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \frac{\pi(\mathbf{A}_i; \alpha) I(\Delta_{ij} = 1) I(X_{ij} \leq t)}{\Pr(\mathbf{A}_i | \mathbf{L}_i, \hat{\boldsymbol{\beta}}) \Pr(\Delta_{ij} = 1 | \mathbf{L}_i, X_{ij}, \hat{\boldsymbol{\gamma}})},$$

and $\hat{\boldsymbol{\gamma}}$ is an estimator of the vector of the parameters for the censoring model. Details of the censoring model are discussed in the next sections. Estimates of the direct, indirect, total, and overall effects are given by $\widehat{DE}(t, \alpha) = \hat{\mu}(t, 0, \alpha) - \hat{\mu}(t, 1, \alpha)$, $\widehat{IE}(t, \alpha_1, \alpha_2) = \hat{\mu}(t, 0, \alpha_1) - \hat{\mu}(t, 0, \alpha_2)$, $\widehat{TE}(t, \alpha_1, \alpha_2) = \hat{\mu}(t, 0, \alpha_1) - \hat{\mu}(t, 1, \alpha_2)$ and $\widehat{OE}(t, \alpha_1, \alpha_2) = \hat{\mu}(t, \alpha_1) - \hat{\mu}(t, \alpha_2)$.

Known Treatment and Censoring Distributions

The proposition below shows that if the group level propensity scores and the individual censoring probabilities are known, then the proposed IPCW estimators are unbiased. A

proof of the proposition is given in Section 2.6.

Proposition 1. *If* $\Pr(\mathbf{A}_i|\mathbf{L}_i)$ *and* $\Pr(\Delta_{ij}=1|\mathbf{L}_i)$ *are known for all* $j=1,2,\ldots,n_i$, *then* $E\{\hat{F}_i(t,a,\alpha)\}=\bar{F}_i(t,a,\alpha)$ *and* $E\{\hat{F}_i(t,\alpha)\}=\bar{F}_i(t,\alpha)$.

Unknown Treatment and Censoring Distributions

In observational studies, neither the conditional distribution of treatment given covariates nor the conditional distribution of censoring given covariates are known. Therefore, we consider finite dimensional parametric models to estimate the group propensity scores and conditional probability of censoring; these estimates are then plugged into the IPCW estimators defined above. The conditional probability of censoring is estimated using a frailty model (Munda et al. 2012) where the conditional hazard for C_{ij} is assumed to have the form $g_{ij}(c|\mathbf{L}_{ij},e_i)=g_0(\boldsymbol{c};\boldsymbol{\theta_h})e_i\exp(\mathbf{L}_{ij}^T\boldsymbol{\theta_c})$, where g_0 is the baseline hazard function, $\boldsymbol{\theta_h}$ is the q'- dimensional parameter vector of the baseline hazard function, e_i is a random effect with density $f_e(e_i;\theta_r)$, and $\boldsymbol{\theta_c}$ is the q-dimensional vector of coefficients. Let $\boldsymbol{\gamma}=(\boldsymbol{\theta_c},\boldsymbol{\theta_h},\theta_r)$ be the vector of parameters for the frailty model. Maximum likelihood theory can be used to draw inference about $\boldsymbol{\gamma}$. Under assumption III, the contribution of group i to the marginal log-likelihood is (Munda et al. 2012)

$$l(\mathbf{X_i},\boldsymbol{\Delta_i},\mathbf{L_i},\boldsymbol{\gamma})=\sum_{j=1}^{n_i}\Delta_{ij}\left[\log\{g_0(X_{ij})\}+\mathbf{L}_{ij}^T\boldsymbol{\theta_c}\right]+(-1)^{d_i}\mathcal{L}^{(d_i)}\sum_{j=1}^{n_i}G_0(X_{ij})\exp(\mathbf{L}_{ij}^T\boldsymbol{\theta_c}),$$

where $d_i=\sum_{j=1}^{n_i}(1-\Delta_{ij})$ is the number of censored observations in group i, $G_0(\omega)=\int_0^\omega g_0(\kappa)d\kappa$, and $\mathcal{L}^{(s)}$ is the s-th derivative of the Laplace transform of the frailty distribution, i.e., $\mathcal{L}^{(s)}=\int_0^\infty\exp(-e_is)f_e(e_i;\theta_r)de_i$. Therefore, the maximum likelihood

estimator of $\boldsymbol{\gamma}$ solves the following estimating equations

$$\sum_i \psi_{ck}(\mathbf{X}_i, \boldsymbol{\Delta}_i, \mathbf{L}_i, \boldsymbol{\gamma}) = 0 \text{ for } k = 1, ..., q + q' + 1,$$

where $\psi_{ck} = \psi_{ck}(\mathbf{X}_i, \boldsymbol{\Delta}_i, \mathbf{L}_i, \boldsymbol{\gamma}) = \partial l(\mathbf{X}_i, \boldsymbol{\Delta}_i, \mathbf{L}_i, \boldsymbol{\gamma})/\partial\gamma_k$ and γ_k is the k-th element of $\boldsymbol{\gamma}$.

Below, the baseline hazard for the censoring model is assumed to be constant and equal to θ_h such that $g_{ij}(c|\mathbf{L}_{ij}, e_i) = \theta_h \exp(\mathbf{L}_{ij}\boldsymbol{\theta_c})e_i$. In addition, the frailty term e_i is assumed to follow a gamma distribution with mean 1 and variance θ_r. So, the censoring weight for an uncensored individual equals

$$\begin{aligned}
\Pr(\Delta_{ij} = 1|\mathbf{L}_i, \boldsymbol{\gamma}) &= \int \Pr(C_{ij} > T_{ij}(\mathbf{A}_i)|\mathbf{L}_i, \boldsymbol{\gamma}, e_i) f_e(e_i; \theta_r) de_i \\
&= \int \Pr(C_{ij} > X_{ij}|\mathbf{L}_i, \boldsymbol{\gamma}, e_i) f_e(e_i; \theta_r) de_i \\
&= \int \exp\{-\theta_h X_{ij} \exp(\mathbf{L}_{ij}\boldsymbol{\theta_c})e_i\} \frac{e_i^{1/\theta_r - 1} e^{-e_i/\theta_r}}{\theta_r^{1/\theta_r} \Gamma 1/\theta_r} de_i \\
&= \left\{ \frac{1}{\theta_r \theta_h X_{ij} \exp(\mathbf{L}_{ij}\boldsymbol{\theta_c}) + 1} \right\}^{1/\theta_r}
\end{aligned}$$

Following TV (2012), a mixed effects model may be assumed for the treatment allocation, i.e., $\Pr(A_{ij} = 1|\mathbf{L}_{ij}, b_i) = \text{logit}^{-1}(\mathbf{L}_{ij}\boldsymbol{\theta_x} + b_i)$ where b_i is a random effect following density $f_b(b_i; \theta_s)$. (In the application below the mixed effects model has a slightly more complicated form owing the particulars of the design of the study analyzed.) Let $\boldsymbol{\beta} = (\boldsymbol{\theta_x}, \theta_s)$ denote the $(p + 1)$ dimensional vector of parameters for the mixed effects model. Again, maximum likelihood theory can be used to draw The contribution of group i to the log-likelihood for the mixed effects model is given by

$$l(\mathbf{A}_i, \mathbf{L}_i, \boldsymbol{\beta}) = \log \left[\int \prod_{j=1}^{n_i} h_{ij}(b_i, \mathbf{L}_i, \boldsymbol{\theta_x})^{A_{ij}} \{1 - h_{ij}(b_i, \mathbf{L}_i, \boldsymbol{\theta_x})\}^{(1-A_{ij})} f_b(b_i; \theta_s) \right],$$

where $h_{ij}(b_i, \mathbf{L}_i, \boldsymbol{\beta}) = \Pr(A_{ij} = 1|\mathbf{L}_{ij}, b_i)$. The maximum likelihood estimator of $\boldsymbol{\beta}$ is the

solution to the score equations

$$\sum_i \psi_{xk}(\mathbf{A}_i, \mathbf{L}_i, \boldsymbol{\beta}) = 0 \text{ for } k = 1, ..., p+1,$$

where $\psi_{xk} = \psi_{xk}(\mathbf{A}_i, \mathbf{L}_i, \boldsymbol{\beta}) = \partial l(\mathbf{A}_i, \mathbf{L}_i, \boldsymbol{\beta})/\partial \beta_k$, β_k is the k-th element of $\boldsymbol{\beta}$.

Inference about the causal effects of interest is then based on solving the vector of estimating equations

$$\sum_i \psi(\mathbf{O}_i, \boldsymbol{\theta}) = 0, \tag{2.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \theta)$, $\psi(\mathbf{O}_i, \boldsymbol{\theta}) = (\boldsymbol{\psi_c}, \boldsymbol{\psi_x}, \psi_{a\alpha})^T$, $\boldsymbol{\psi_c} = (\psi_{c1}, \psi_{c2}, ..., \psi_{cq+q'+1})^T$, $\boldsymbol{\psi_x} = (\psi_{x1}, \psi_{x2}, ..., \psi_{xp+1})^T$,

$$\psi_{a\alpha} = \psi_{a\alpha}(\mathbf{O}_i, \boldsymbol{\theta}) = \frac{g^*(\mathbf{O}_i, a, \alpha, \boldsymbol{\gamma})}{\Pr(\mathbf{A}_i|\mathbf{L}_i, \boldsymbol{\beta})} - \theta,$$

and

$$g^*(\mathbf{O}_i, a, \alpha, \boldsymbol{\gamma}) = n_i^{-1} \sum_{j=1}^{n} \frac{\pi(\mathbf{A}_{i,-j}; \alpha) I(A_{ij} = a) I(X_{ij} \leq t)}{\Pr(\Delta_{ij} = 1|\mathbf{L}_{ij}, X_{ij}, \boldsymbol{\gamma})}.$$

Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}, \hat{\mu}(t, a, \alpha))$ denote the solution to (1). Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta_0} = (\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0, \mu(t, a, \alpha))$ and note that

$$\int \psi_{a\alpha}(\mathbf{o}, \boldsymbol{\gamma}_0, \boldsymbol{\beta}_0, \mu(t, a, \alpha)) dF_{\mathbf{O}}(\mathbf{o}) = E\left\{ \frac{g^*(\mathbf{O}_i, a, \alpha, \boldsymbol{\gamma}_0)}{\Pr(\mathbf{A}_i|\mathbf{L}_i, \boldsymbol{\beta}_0)} - \mu(t, a, \alpha) \right\} = 0,$$

where $F_{\mathbf{O}}$ denotes the joint distribution of the complete observed random variable $\mathbf{O}$ and the last equality follow from the Proposition above. Therefore, assuming the parametric models above are correctly specified, it follows that $\int \psi(\mathbf{o}, \boldsymbol{\theta}_0) dF_O(\mathbf{o}) = 0$. By M-estimation theory (Stefanski and Boos 2002), $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta_0}$ and $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})$ converges in distribution to a Normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}$ equal to $U(\boldsymbol{\theta}_0)^{-1} V(\boldsymbol{\theta}_0) \{U(\boldsymbol{\theta}_0)^{-1}\}^T$ where $U(\boldsymbol{\theta}_0) = E\{-\dot{\psi}(\mathbf{O}_i, \boldsymbol{\theta_0})\}$, $V(\boldsymbol{\theta}_0) = E\{\psi(\mathbf{O}_i, \boldsymbol{\theta_0}) \psi(\mathbf{O}_i, \boldsymbol{\theta_0})^T\}$, and $\dot{\psi}(\mathbf{O}_i, \boldsymbol{\theta}) = \partial \psi(\mathbf{O}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}^T$. Consistency and asymptotic

normality of the direct, indirect and total effect estimators follows from the delta method. Similar techniques can be used to show that $\hat{\mu}(t, \alpha)$ and the overall effect estimator are also consistent and asymptotically Normal. The asymptotic variance $\boldsymbol{\Sigma}$ can be consistently estimated by $\hat{\boldsymbol{\Sigma}} = \hat{U}(\hat{\boldsymbol{\theta}})^{-1}\hat{V}(\hat{\boldsymbol{\theta}})\{\hat{U}(\hat{\boldsymbol{\theta}})^{-1}\}^T$ where $\hat{U}(\hat{\boldsymbol{\theta}}) = m^{-1}\sum_{i=1}^{m}\{-\dot{\psi}(\mathbf{O}_i, \hat{\boldsymbol{\theta}})\}$ and $\hat{V}(\hat{\boldsymbol{\theta}}) = m^{-1}\sum_{i=1}^{m}\{\psi(\mathbf{O}_i, \hat{\boldsymbol{\theta}})\psi(\mathbf{O}_i, \hat{\boldsymbol{\theta}})^T\}$. The empirical sandwich variance estimator $\hat{\boldsymbol{\Sigma}}$ can be computed using the R package `geex` (Saul and Hudgens 2017) and can be used to construct Wald type confidence intervals (CIs).

2.3 Simulation Study

A simulation study was conducted to assess the finite sample bias of the IPCW estimator and coverage of the corresponding Wald confidence intervals. The data generating model used in the simulation study was motivated by aspects of the cholera vaccine study analysis presented in the next section. Following Perez-Heydrich et al. (2014), data were simulated according to the following steps.

i) First, two baseline covariates L_{1ij} and L_{2ij} were randomly generated. In the application presented in Section 2.4, conditional exchangeability is assumed given an individual's age (in decades) and the distance of their residence to the nearest river. Motivated by this example, L_{1ij} and L_{2ij} were randomly generated as follows. First, V_{ij} was randomly generated from an exponential distribution with mean 20. Then L_{1ij} was set to $\min(V_{ij}, 100)/10$. The second set of covariates L_{2ij} were randomly sampled such that $\log L_{2ij} \sim \text{normal}(0, 0.75)$.

ii) The random effects for the treatment model b_i were randomly sampled from a normal distribution with mean 0 and variance 0.0859.

iii) The treatment indicators A_{ij} were randomly sampled from a Bernoulli distribution with mean $p_{ij} = \text{expit}(0.2727 - 0.0387L_{1ij} + 0.2179L_{2ij} + b_i)$.

iv) The potential times to event $T_{ij}(\mathbf{a}_i)$ were randomly sampled from an exponential distribution with mean $\mu_{ij} = 200 + 100a_{ij} - 0.98L_{1ij} - 0.145L_{2ij} + 50\sum_{k\neq j} a_{ik}/n_i$.

v) The random effects for the censoring model e_i were randomly generated from a gamma distribution with mean 1 and variance $\theta = 1.25$.

vi) Censoring times C_{ij} were randomly sampled from an exponential distribution with mean $1/\lambda_0$ where $\lambda_0 = 0.015\exp(0.002L_{1ij} + 0.015L_{2ij})e_i$.

vii) Individual censoring indicators were determined i.e., $\Delta_{ij} = 0$ if $C_{ij} < T_{ij}(\mathbf{A}_i)$.

Steps i through vii were used to stochastically generate 1000 data sets, with each data set containing 500 groups with 10 individuals per group. For each simulated data set, the IPCW estimator of $\mu(100, a, \alpha)$ was evaluated for $a = 0, 1$ and $\alpha = 0.1, 0.2, \ldots, 0.9$. Estimated standard errors based on the empirical sandwich variance estimator and Wald 95% confidence intervals were also calculated for each simulated data set. Empirical standard errors were calculated by taking the standard deviation of the point estimates from all simulations.

The true value of the estimand was obtained by simulating counterfactual outcomes for $m = 10^6$ groups of individuals. Note that, according to the model used to generate the data, potential survival times depend only on $\sum_{k\neq j} a_{ik}$. So, $\mu(t, a, \alpha)$ was approximated by (Perez-Heydrich et al. 2014)

$$m^{-1}\sum_{i=1}^{m} n_i^{-1}\sum_{j=1}^{n_i}\sum_{k=0}^{n_i-1}\binom{n_i-1}{k} I\{T_{ij}(a,k) \leq t\}\alpha^k(1-\alpha)^{n_i-k-1}.$$

The true value of $\mu(t, \alpha)$ was determined in a similar fashion.

Results from the simulation study are presented in Table 2.1. Bias of the IPCW estimator was negligible for all values of a and α. Likewise, the average estimated standard error was close to the empirical standard error. Coverage of the 95% Wald CIs was approximately equal to the nominal level.

Additional simulation studies were conducted to assess the performance of the proposed methods for different values of m, the total number of groups, ranging from 10 to 500. The number of individuals per group was 10, as in the previous simulations. For each $m \in \{10, 50, 100, 200, 300, 400, 500\}$, 1000 data sets were simulated according to steps i through vii above. Results are depicted in Figure 2.1. Bias of the IPCW estimator was small and coverage of the Wald CIs was close to the nominal level provided m was at least 50.

2.4 Data Analysis

2.4.1 Cholera Vaccine Study and Analysis

In this section, the methods described in Section 2.2 are used to analyze a cholera vaccine study in Matlab, Bangladesh (Ali et al. 2005). Eligible study participants were children 2–15 years of age and women greater than 15 years old. All $121,975$ eligible individuals in the population were randomized to one of three vaccination groups: B subunit-killed whole-cell oral cholera vaccine, killed whole-cell-only cholera vaccine, and *E. coli* K12 placebo. As in Perez-Heydrich et al. (2014), no distinction is made between the two vaccines in the analysis presented here. Individuals were considered to have participated in the randomized trial component of the study if they received two or more doses of vaccine or placebo. The primary endpoint of the trial was incident cholera. Three health centers in the Matlab area served as surveillance centers and collected endpoint data on all individuals, regardless of whether they participated in the randomized trial. The analysis presented here includes data from all individuals, i.e., trial participants as well as those who chose not to participate. Thus an approach which accounts for possible confounding, such as the IPW method described in Section 2, should be utilized to assess the effects of vaccination.

Previous analyses of this study suggest the presence of interference (Ali et al. 2005, Perez-Heydrich et al. 2014). However, these previous analyses did not formally account

for censoring. Here individuals are considered right censored if they were not diagnosed with cholera during the study. Individuals who emigrated from the study location or died during the follow-up period prior to cholera infection were right censored at the time of emigration or death. Individuals who did not emigrate or die and who did not develop cholera during the study were right censored at the end of the study period.

Related individuals in Matlab live in clustered sets of houses called baris. There were a total of 6,415 baris at the time of the vaccine trial. Perez-Heydrich et al. (2014) used a clustering algorithm to form groups (neighborhoods) based on the spatial location of the baris, with the number of groups pre-specified to be 700. The analysis here is based on the same groups as in Perez-Heydrich et al. and assumes that there is no interference between individuals in different groups, i.e, the vaccination of an individual in one group has no effect on whether an individual in another group acquires cholera. When fitting the propensity model $\Pr(\mathbf{A}_i|\mathbf{L}_i, \boldsymbol{\beta})$ described below, the largest 15 groups had estimated group propensity scores that were effectively equal to zero and therefore these groups were omitted.

Individuals participating in the vaccine trial were not all vaccinated on the same calendar day, such that the level of vaccine coverage within a group varied over a relatively brief period of calendar time at the study onset. For simplicity and because the methods developed above do not accommodate time varying treatment, the start of follow-up for all individuals in a particular group was set to the latest date of second vaccination among all individuals in that group. Some observations were excluded because individuals contracted cholera, died, or emigrated prior to the start of follow-up for their group.

In total, 94,234 individuals were included in the analysis. Among these individuals, 55,413 were unvaccinated, either because they received placebo or they did not participate, and 38,821 were vaccinated with one of the two vaccines. During follow-up, there were 280 incident cases of cholera among the unvaccinated individuals and 74 cholera cases among the vaccinated individuals.

As in Perez-Heydrich et al., the group propensity score was modeled using a mixed effects model. The particular form of the model derives from the fact that in order for an individual to have received a vaccine, they must have (i) chosen to participate in the trial, and (ii) been randomized to receive one of the two vaccines. To account for (i), a logistic regression model for participation was assumed. As in Perez-Heydrich et al., covariates in the participation component of the model were age, squared age, distance to nearest river, and squared distance to nearest river. Accommodating (ii) in the propensity model is straightforward because, due to randomization, individuals who elected to participate in the trial were known to receive one of the two vaccines with probability 2/3. Combining these two aspects of the model, the propensity score for group i was estimated by

$$\Pr(\mathbf{A}_i|\mathbf{L}_i,\hat{\boldsymbol{\beta}}) = \int \prod_{j=1}^{n_i}\{(2/3)h_{ij}(b_i,\mathbf{L}_{ij},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\}^{A_{ij}}\{1-(2/3)h_{ij}(b_i,\mathbf{L}_{ij},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\}^{(1-A_{ij})} \times f_b(b_i;\hat{\theta}_s),$$

where $h_{ij}(b_i,\mathbf{L}_i,\boldsymbol{\theta}_x) = \Pr(B_{ij}=1|b_i,\mathbf{L}_{ij},\boldsymbol{\theta}_x) = \text{expit}(\mathbf{L}_{ij}\boldsymbol{\theta}_{\boldsymbol{x}} + b_i)$, B_{ij} is the indicator of participation, i.e., $B_{ij}=1$ if individual j in group i participated in the randomized trial and $B_{ij}=0$ otherwise, and $(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}},\hat{\theta}_s)$ is the maximum likelihood estimate of $(\boldsymbol{\theta}_{\boldsymbol{x}},\theta_s)$. Censoring was modeled using the gamma frailty model described above, and only included age as covariate as no other variables were associated with censoring. Over 70% of individuals belonged to groups where the vaccine coverage was between 0.3 and 0.6. Therefore, the analysis was conducted for allocation strategies ranging from 0.3 to 0.6.

2.4.2 Results

Figure 2.2 shows the IPCW estimates of the cumulative probability of cholera over time for allocation strategies 0.3, 0.45, and 0.6, both when an individual receives a vaccine and when an individual is unvaccinated. The estimated risk of cholera when an individual is unvaccinated decreases dramatically as α increases, suggesting the presence

of interference. This decrease is more modest when an individual is vaccinated, indicating a stronger indirect effect when unvaccinated. At all time points the estimated risk of cholera is higher when an individual is unvaccinated, suggesting a beneficial, direct effect of vaccination, especially at lower coverage levels. For $\alpha = 0.3$ and $\alpha = 0.45$, the estimated risk when unvaccinated increases suddenly between 200 and 300 days, and then again between 300 and 400 days. These results might be attributable to the known bimodal seasonality of cholera in Bangladesh (Longini et al. 2002). Note that, because the study start date varied across groups, the time scale in this analysis does not exactly coincide with calendar time. Nonetheless, 95% of individuals had a start date within a two calendar month range, such that there is a strong correlation between the analysis time scale and calendar time, and thus cholera seasonality may explain these periods of marked increase in risk.

Direct, indirect, total and overall effect estimates and 95% CIs ($\times 1000$) for different allocation strategies at time $t = 1$ year are shown in Figure 2.3. The direct effect estimates generally decrease as α increases. For example, the direct effect estimate for $\alpha = 0.35$ is 3.6 (95% CI $1.1, 6.2$) whereas for $\alpha = 0.5$ the direct effect estimate is 1.5 (95% CI $-0.5, 3.5$). The indirect, total, and overall effect estimates in Figure 2.3 compare the risk of cholera over a range of allocation probabilities $\alpha_1 \in [0.3, 0.6]$ versus $\alpha_2 = 0.4$. Here the indirect effect contrasts risk of cholera infection when individuals are unvaccinated. For larger values of α_1 the 95% CIs for these effects exclude the null value of zero. For example, for $\alpha_1 = 0.6$ the indirect effect estimate is 2.8 (95% CI 1.1, 4.5), providing statistical evidence of the presence of interference. These results indicate that when individuals are unvaccinated, the risk of cholera infection is significantly reduced by increasing the level of vaccine coverage in their neighborhood. The total effect estimates quantify the combined direct and indirect effects of the vaccine. The overall effect estimates may be of greatest interest from a public health or policy perspective. For $\alpha_1 = 0.6$, the overall effect estimate is 2.2 (95% CI 0.9, 3.4); in words, 2.2 fewer cases of

cholera per 1000 individuals per year are expected if 60% of individuals are vaccinated compared to if only 40% of individuals receive vaccine.

In previous analyses of these data, Perez-Heydrich et al. also estimated the direct, indirect, total and overall effects using a binary outcome indicating whether an individual was infected with cholera during the first year of follow-up. The IPCW estimates for $t = 1$ are similar to these previous results, e.g., Perez-Heydrich et al. estimated the direct effect for alpha=0.32 to be 5.3 (95% CI 2.5, 8.1) whereas the IPCW estimate of this effect at $t = 1$ is 4.0 (95% CI 1.6, 6.5). However,the Perez-Heydrich et al. estimates may be biased because they did not account for right censoring.

2.5 Discussion

In this section, the TV IPW estimator for partial interference was extended to allow for right censored outcomes. The proposed estimator was obtained by weighting the original TV estimator by censoring weights calculated from a parametric frailty model of the censoring times. The estimator was shown to be consistent and asymptotically normal and a consistent estimator of the asymptotic variance was proposed. A simulation study demonstrated that the proposed methods performed well in finite samples provided the number of groups is sufficiently large. Analysis of a cholera vaccine study using the proposed methods suggests vaccination had both a direct and indirect effect against cholera infection. These results are in accordance with findings by Ali et al. (2005) and Perez-Heydrich et al. (2014), but are likely more accurate since these previous analyses did not formally account for right censoring.

There are several areas of possible future research related to the methods developed here. For example, further research could entail developing estimators which perform well in settings where the number of groups is small. Alternative IPCW estimators could be developed which utilize semi-parametric frailty models to estimate the censoring weights rather than the fully-parametric models employed here. Extensions of the IPCW

estimator could also be considered for the setting where there is general interference, i.e., where interference is not restricted to individuals within the same group. In this paper only Horwitz-Thompson type IPCW estimators were considered; further research could entail developing stabilized or Hajek type IPCW estimators which may be more stable and less variable. Finally, simulations studies (results not shown here) suggest that the proposed IPCW estimators may be sensitive to model mis-specification. Future research could entail developing estimators that are robust to model mis-specification, perhaps by constructing doubly robust estimators which utilize both a treatment model and an outcome model.

Figure 2.1: Absolute bias (left) and 95% confidence interval coverage (right) for different numbers of groups for $\alpha = 0.5$. The dotted line in the right plot corresponds to 95% coverage.

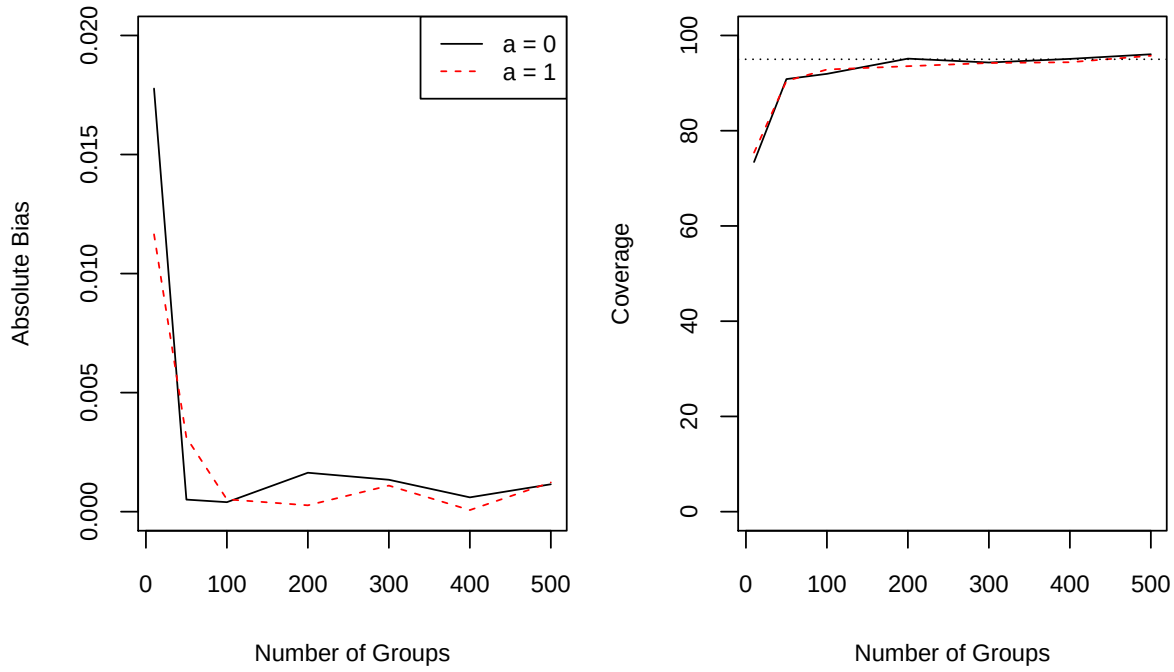


Figure 2.2: Estimated cumulative probability of cholera over time for vaccinated and unvaccinated for $\alpha = 0.3$ (left), $\alpha = 0.45$ (center) and $\alpha = 0.6$ (right)

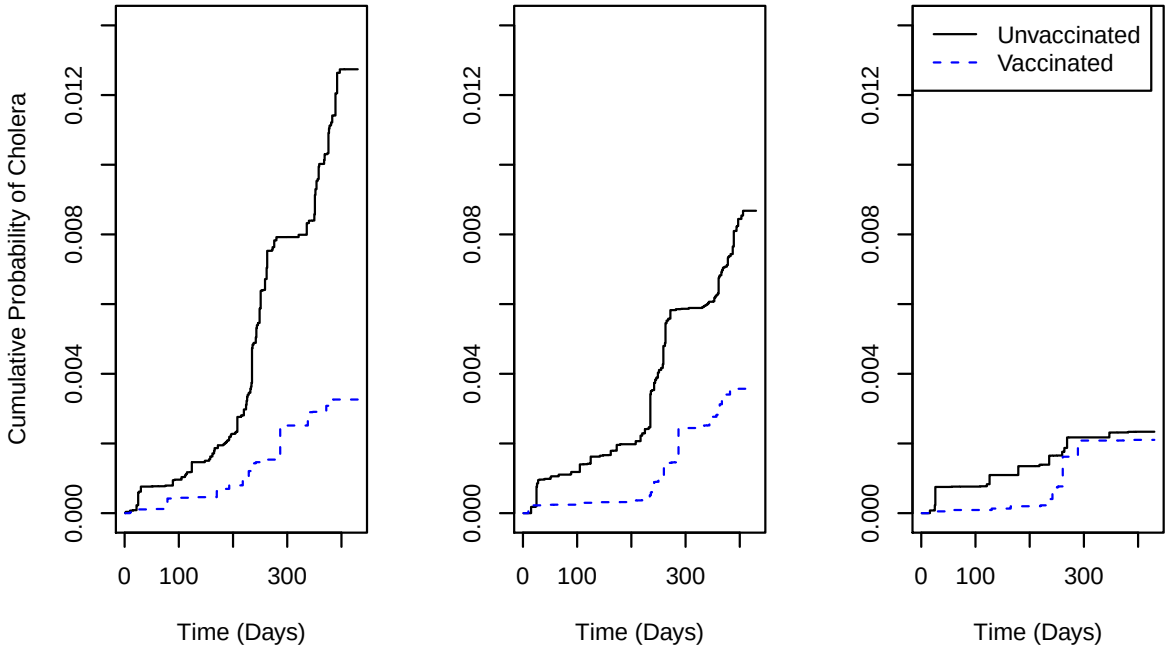
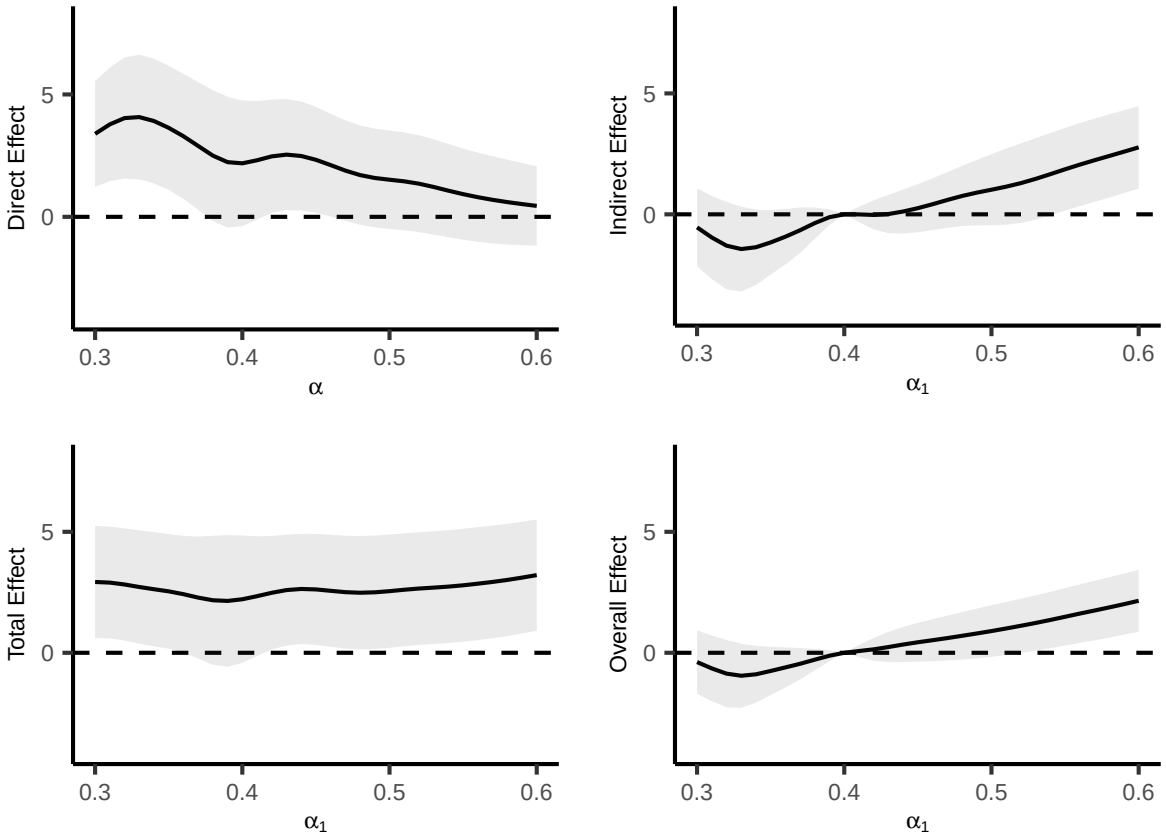


Figure 2.3: Direct, indirect, total and overall effect estimates ($\times 1000$) for different allocation strategies at time $t = 1$ year. Indirect, total, and overall effects are with respect to $\alpha_2 = 0.4$. The shaded regions denote pointwise 95% confidence intervals of the estimates.



α	$\mu(100,0,\alpha)$	Bias	ESE	ASE	EC	α	$\mu(100,1,\alpha)$	Bias	ESE	ASE	EC
0.1	0.39	0.02	0.07	0.07	94%	0.1	0.28	0.01	0.08	0.08	92%
0.2	0.38	0.01	0.04	0.04	96%	0.2	0.27	0.01	0.04	0.04	95%
0.3	0.38	0.00	0.03	0.03	96%	0.3	0.27	-0.00	0.03	0.03	95%
0.4	0.37	-0.00	0.03	0.02	95%	0.4	0.27	-0.01	0.02	0.02	94%
0.5	0.36	-0.00	0.03	0.02	94%	0.5	0.26	-0.00	0.02	0.02	93%
0.6	0.36	-0.01	0.03	0.02	94%	0.6	0.26	-0.00	0.02	0.02	93%
0.7	0.35	-0.00	0.03	0.02	94%	0.7	0.26	-0.01	0.02	0.01	94%
0.8	0.35	-0.01	0.03	0.03	94%	0.8	0.25	-0.00	0.02	0.02	93%
0.9	0.34	-0.00	0.05	0.05	92%	0.9	0.25	0.01	0.02	0.02	95%

Table 2.1: Results from simulation study described in Section 2.3. α denotes the allocation probability, $\mu(100, a, \alpha)$ is the true value of the target parameter for $a = 0, 1$; Bias is the average of $\mu(100, a, \alpha) - \hat{\mu}(100, a, \alpha)$ for $a = 0, 1$; ESE is the empirical standard error; ASE is the average of the sandwich variance estimates; and EC denotes the empirical coverage of the 95% Wald confidence intervals.

CHAPTER 3: PARAMETRIC G-FORMULA WITH PARTIAL INTERFERENCE AND RIGHT CENSORING

3.1 Introduction

Interference is present when the outcome of one individual depends on the treatment status of another individual (Cox 1958). An example of this might be data on infectious diseases. For these types of data, whether a subject becomes infected or not might be affected by the vaccination status of another individual (Halloran and Struchiner 1991). Sobel (2006) introduces the notion of partial interference as a subset of interference. If individuals can be partitioned into groups such that interference can occur within individuals of one group but it cannot occur between individuals from two separate groups, the data is said to show traits of partial interference. Well defined social, temporal, and/or geographical difference between groups of people might be a valid reason for assuming they have partial interference. Interference might produce effects termed spillover effects or peer effects, which are of importance for various different fields of study.

There are inference methods available for randomized experiments in the presence of interference (e.g., Rosenbaum 2007, Hudgens and Halloran 2008, Baird et al 2014, Eckles et al 2016). But it might not always be possible to conduct randomized experiments due to feasibility and/or ethical issues. Tchetgen Tchetgen and VanderWeele (TV) (2012) provided consistent estimates for various causal effects in the absence of randomization i.e. for observational data using inverse probability weighting. However, there are some significant disadvantages of using IPW estimators. For example, a propensity score close to zero might make the estimator unstable and difficult to calculate computationally. A substitute of the inverse probability weighted estimator to calculate causal effects is to use

the parametric g formula. The generalized computation algorithm formula (g formula) was first put forward by Robins (1986). G-formula, together with outcome regression produces the parametric g formula. The theory of parametric g formula is generalized from standardization (Hernán and Robins 2006). Parametric g formula has been used mainly in adjusting for time varying confounders for time to event data (Young et al. (2011)). Many authors used parametric g formula to calculate risk ratio (Taubman et al. (2009), Garcia-Aymerich et al. (2013), Cole et al. (2013)) and hazard ratio (Westreich et al. (2012), Keil et al. (2014)). However, all of the authors have used logistic regression for modeling the probability of outcome.

The rest of the section is organized as follows. Section 3.2 introduces the estimators and estimands. Results from simulation studies illustrating the finite sample behavior of the method are presented in Section 3.3. The proposed method is implemented on a real data consisting of 100,000 individuals in Matlab, Bangladesh in Section 3.4.

3.2 Methods

3.2.1 Estimands

Assume that there are m groups, each group having n_i individuals in them for $i = 1, \ldots, m$. Depending on whether individual j in group i gets treatment or placebo, denote $A_{ij} = 1$ or $A_{ij} = 0$ respectively. Let $\mathbf{A}_i = (A_{i1}, A_{i2}, ..., A_{in_i})$ and $\mathbf{A}_{i,-j} = (A_{i1}, A_{i2}, ..., A_{ij-1}, A_{ij+1}..., A_{in_i})$ denote the overall group treatment assignment and the group treatment assignment without individual j for group i respectively. Also, let $\mathbf{A}_i$ and $\mathbf{A}_{i,-j}$ attain possible values $\mathbf{a}_i$ and $\mathbf{a}_{i,-j}$ respectively. The potential time to event for individual j in group i with treatment $\mathbf{a}_i$ is denoted by $T_{ij}(\mathbf{a}_i)$. These potential times exhibit traits of partial interference in the sense that the potential time to event $T_{ij}(\mathbf{a}_i)$ for individual j in group i might be dependent on the treatment status of individual j' in group i even when j' does not equal j. However, if $i \neq i'$, then $T_{ij}(\mathbf{a}_i)$ is independent of $A_{i'j}$ for any j. The set of all potential event times for individuals in group

i is denoted by $\mathbf{T}_i(.) = \{T_{ij}(\mathbf{a}_i) : \mathbf{a}_i \in \mathcal{A}(n_i), j = 1, 2, \cdots, n_i\}$. Assume that there is right censoring within the time to events. Also assume that C_{ij} are censoring time for individual j in group i. Let $\Delta_{ij} = I(T_{ij}(\mathbf{A}_i) \leq C_{ij})$ and $X_{ij} = \min(T_{ij}(\mathbf{A}_i), C_{ij})$. Then Δ_{ij} is the censoring indicator and X_{ij} is the observed time to event for individual j in group i. The vector of censoring indicators for group i denoted by $\mathbf{\Delta}_i$ equals $(\Delta_{i1}, \Delta_{i2}, \cdots, \Delta_{in_i})$ and the vector of observed time to events denoted by $\mathbf{X}_i$ equals $(X_{i1}, X_{i2}, \cdots, X_{in_i})$. The set of every feasible 2^n assignments of treatments for $n = 1, 2, \ldots$ is denoted by $\mathcal{A}(n)$. Finally, the vector of all the covariates for subject j in group i is termed as $\mathbf{L}_{ij}$ and $\mathbf{L}_i = (\mathbf{L}_{i1}, \mathbf{L}_{i2}, \cdots, \mathbf{L}_{in_i})$ denotes the matrix of covariates for group i. Let the baseline covariates $\mathbf{L}_{ij}$ include group sizes n_i. The m groups in data is assumed to be sampled from an infinite superpopulation of groups and the m observations $(\mathbf{L}_i, \mathbf{A}_i, \mathbf{X}_i, \mathbf{\Delta}_i)$ are i.i.d..

When there is no possible interference, there is often interest in finding the average treatment effect. This can be represented as the difference between two counterfactual outcomes, the first one being the case where all the individuals in the population are treated and the second one being the case when no individual in the population are treated. But if interference is present in a data then the group allocation strategies α might affect the counterfactual outcomes of interest (Hong and Raudenbush 2006, Sobel 2006, Tchetgen and VanderWeele 2012, Hudgens and Halloran 2008). Tchetgen Tchetgen and Vanderweele (2012) discussed how α can be explained as a "Bernoulli" treatment allocation strategy. A possible interpretation of α might be the probability of being assigned treatment independent of others for an individual. Hence, a natural extrapolation of the concept of average treatment effect might be calculating the difference between the counterfactual outcomes corresponding to two different levels of allocation strategies α and α'. Again, let the conditional probability that the treatment assignment for group i except for individual j is $\mathbf{a}_{i,-j}$ given that the j^{th} individual in the i^{th} group

receives treatment a under allocation strategy α be denoted by $\pi(\mathbf{a}_{i,-j}, \alpha)$. Mathematically, this can be denoted as follows $\pi(\mathbf{a}_{i,-j}, \alpha) = \Pr(\mathbf{A}_{i,-j} = \mathbf{a}_{i,-j} | A_{i,j} = a)$. Then, $\pi(\mathbf{a}_{i,-j}, \alpha) = \Pi_{k=1,k\neq j}^{n_i} \alpha^{a_{ik}}(1-\alpha)^{1-a_{ik}}$. Similarly, let $\pi(\mathbf{a}_i, \alpha)$ denote the conditional probability that the treatment assignment for the i^{th} group is $\mathbf{a}_i$ under allocation strategy α. In terms of probability, $\pi(\mathbf{a}_i, \alpha) = \Pr(\mathbf{A}_i = \mathbf{a}_i)$. Then, according to the independent Bernoulli probability assumption for α, $\pi(\mathbf{a}_i, \alpha) = \Pi_{k=1}^{n_i} \alpha^{a_{ik}}(1-\alpha)^{1-a_{ik}}$.

The various contrasts in survival probabilities for different combinations of treatment and allocation strategies are possible causal estimands of interest. To properly introduce these estimands, first the following has to be defined,

$$\bar{F}_{ij}(t, a, \alpha) = \sum_{a_{i,-j} \in \mathcal{A}(n_i - 1)} I\{T_{ij}(a, \mathbf{a}_{i,-j}) \leq t\} \pi(\mathbf{a}_{i,-j}, \alpha),$$

$$\bar{F}_{ij}(t, \alpha) = \sum_{a_i \in \mathcal{A}(n_i)} I\{T_{ij}(\mathbf{a}_i) \leq t\} \pi(\mathbf{a}_i, \alpha),$$

$\bar{F}_i(t, a, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \bar{F}_{ij}(t, a, \alpha)$ and $\bar{F}_i(t, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \bar{F}_{ij}(t, \alpha)$. $\bar{F}_i(t, a, \alpha)$ can be interpreted as the average probability that an individual will fail by time t in group i when the individual receives a and the group adopts allocation strategy α. Similarly, $\bar{F}_i(t, \alpha)$ can be interpreted as the average probability that an individual will fail by time t in group i when the group adopts allocation strategy α . Finally, $\mu(t, a, \alpha) = E\{\bar{F}_i(t, a, \alpha)\}$ is the population average potential cumulative distribution at the point t for treatment a under allocation strategy α. Similarly $\mu(t, \alpha) = E\{\bar{F}_i(t, \alpha)\}$ is the population average overall potential cumulative distribution at the point t under allocation strategy α. A possible interpretation of $\mu(t, a, \alpha)$ might be as the probability of an individual's survival time being less than t under the counterfactual case that an individual receives treatment a under allocation strategy α,. Again, a possible interpretation of $\mu(t, \alpha)$ might be as the probability of an individual's survival time being less than t under the counterfactual case that the group allocation probability of treatment is α.

The various causal effects of interest are as follows. Comparison of the probability of event of treated individuals with untreated at a particular allocation level gives rise to direct effect. The population average direct effect at time t with the data having allocation probability α is given by $DE(t,\alpha) = \mu(t,0,\alpha) - \mu(t,1,\alpha)$. Subtraction of the probability of event for two different levels of allocation among the untreated gives rise to the indirect effect. For allocation strategies α_1 and α_2, the population average indirect effect is then given by $IE(t,\alpha_1,\alpha_2) = \mu(t,0,\alpha_1) - \mu(t,0,\alpha_2)$. The difference $TE(t,\alpha_1,\alpha_2) = \mu(t,0,\alpha_1) - \mu(t,1,\alpha_2)$ is termed as the total effect. This is the difference between the probability of event of untreated individual at allocation level α_1 and treated individuals at allocation level α_2. The final effect of interest to be discussed is the overall effect. The calculation of this involves subtracting the probability of event of individuals at allocation level α_1 and individuals at allocation level α_2, i.e. $OE(t,\alpha_1,\alpha_2) = \mu(t,\alpha_1) - \mu(t,\alpha_2)$.

3.2.2 Assumptions

The following assumptions are made:

I) Conditional independence: $\mathbf{A}_i \perp\!\!\!\perp \mathbf{T}_i(.)|\mathbf{L}_i$,

II) Positivity: $\Pr(\mathbf{A}_i = \mathbf{a}_i|\mathbf{L}_i = \mathbf{l}) > 0$ for all $\mathbf{a}_i \in \mathcal{A}(n_i)$ and $\mathbf{l}$ such that $\Pr(\mathbf{L}_i = \mathbf{l}) > 0$,

III) Conditional independent censoring: $\mathbf{C}_i \perp\!\!\!\perp \{\mathbf{T}_i(.), \mathbf{A}_i\}|\mathbf{L}_i$.

In the absence of interference, assumption I is a standard assumption made for each individual. When interference is present, this assumption is extended for the group instead of individuals. Under the no interference assumption, assumption (I) is referred to as no unmeasured confounding. In this case, assumption (I) states that the observed treatment assignment for a particular group of individuals is conditionally independent of the counterfactual outcomes of individuals in that group under the possible group treatment statuses under the condition that there is no other confounders except those

observed. Assumption II is termed as positivity. Given all possible values of the measured confounders, the positivity assumption states that given all possible values of the measured confounders, the probability of a group being assigned a particular vector of treatment combination is always positive. Assumption III is related to the censoring distribution. This assumption proposes that the censoring time for an individual is conditionally independent of the counterfactual time to event under and the actual treatment assignment for the group when all the possible confounders are given.

3.2.3 Proposed Estimator

In the absence of interference and censoring, Hernán and Robins (2006) showed that standardization can be used to estimate risk ratios. The form of the standardized estimate of the counterfactual mean $E(Y^a)$ in this case is

$$m(a) = \int \hat{E}(Y|A = a, \mathbf{L} = \mathbf{l}) d\hat{F}_{\mathbf{L}}(\mathbf{l})$$

or equivalently,

$$m(a) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{E}(Y|A = a, \mathbf{L_{ij}})$$

where Y denotes the indicator that an individual has had an event till a specific time point of concern and $\hat{F}_{\mathbf{L}}$ denotes the empirical joint distribution of the covariates $\mathbf{L}$. In previous literature, logistic regression has often been used to estimate $\hat{E}(Y|A = a, \mathbf{L_{ij}})$ (Taubman et al. 2009) and authors had mainly focused on calculating risk ratios or hazard ratios. However, the parametric g formula might be extended for cases with time to event data and interference. The proposed estimator for the causal and survival quantities of interest $\mu(t, a, \alpha)$ defined in Section 3.2.1 is as follows-

$$m^{int}(t, a, \alpha, \hat{\boldsymbol{\omega}}) = \int \sum_{\mathbf{a}_{i,-k} \in \mathcal{A}(n_i - 1)} \Pr(T_{ij} \leq t | \mathbf{L} = \mathbf{l}, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}}) dF_{\mathbf{L}}(\mathbf{l})$$

$$\times \pi(\mathbf{a}_{i,-k}, \alpha),$$

where $\hat{\boldsymbol{\omega}}$ is the estimated value of the parameter for the outcome model $\Pr(T \leq t | \mathbf{L} = \mathbf{l}, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \boldsymbol{\omega})$. The positivity assumption ensures that the probability $\Pr(T \leq t | \mathbf{L} = \mathbf{l}, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \boldsymbol{\omega})$ is well defined. Empirically, the estimator is given by

$$\begin{aligned} m^{int}(t, a, \alpha, \hat{\boldsymbol{\omega}}) = \quad & \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{\mathbf{a}_{i,-j} \in \mathcal{A}(n_i - 1)} \Pr(T_{ij} \leq t | \mathbf{L}_i, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}}) \\ & \times \pi(\mathbf{a}_{i,-k}, \alpha) \\ = \quad & \frac{1}{m} \sum_{i=1}^{m} m_i^{int}(t, a, \alpha, \hat{\boldsymbol{\omega}}) \end{aligned}$$

Similarly, the proposed estimator for the causal and survival quantities of interest $\mu(t, \alpha)$ defined in Section 3.2.1 is as follows-

$$m^{int}(t, \alpha, \hat{\boldsymbol{\omega}}) = \int \sum_{\mathbf{a}_i \in \mathcal{A}(n_i)} \Pr(T_{ij} \leq t | \mathbf{L} = \mathbf{l}, \mathbf{A}_i = \mathbf{a}_i, \hat{\boldsymbol{\omega}}) dF_{\mathbf{L}}(\mathbf{l}) \pi(\mathbf{a}_i, \alpha)$$

Empirically, the estimate is given by

$$\begin{aligned} m^{int}(t, \alpha, \hat{\boldsymbol{\omega}}) &= \frac{1}{m} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{\mathbf{a}_i \in \mathcal{A}(n_i)} \Pr(T_{ij} \leq t | \mathbf{L}_i, \mathbf{A}_i = \mathbf{a}_i, \hat{\boldsymbol{\omega}}) \pi(\mathbf{a}_i, \alpha) \\ &= \frac{1}{m} \sum_{i=1}^{m} m_i^{int}(t, \alpha, \hat{\boldsymbol{\omega}}) \end{aligned}$$

The estimators for $\Pr(T_{ij} \leq t | \mathbf{L} = \mathbf{l}, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \boldsymbol{\omega})$ and $\Pr(T_{ij} \leq t | \mathbf{L} = \mathbf{l}, \mathbf{A}_i = \mathbf{a}_i, \boldsymbol{\omega})$ can be obtained by fitting parametric models. The estimates of the different effects can also be obtained from this. The estimates of the direct, indirect, total and overall effects at a particular time point t are given by $\widehat{DE}(t, \alpha) = m^{int}(t, 0, \alpha, \hat{\boldsymbol{\omega}}) - m^{int}(t, 1, \alpha, \hat{\boldsymbol{\omega}})$, $\widehat{IE}(t, \alpha_1, \alpha_2) = m^{int}(t, 0, \alpha_1, \hat{\boldsymbol{\omega}}) - m^{int}(t, 0, \alpha_2, \hat{\boldsymbol{\omega}})$, $\widehat{TE}(t, \alpha_1, \alpha_2) = m^{int}(t, 0, \alpha_1, \hat{\boldsymbol{\omega}}) - m^{int}(t, 1, \alpha_2, \hat{\boldsymbol{\omega}})$ and $\widehat{OE}(t, \alpha_1, \alpha_2) = m^{int}(t, \alpha_1, \hat{\boldsymbol{\omega}}) - m^{int}(t, \alpha_2, \hat{\boldsymbol{\omega}})$ respectively.

Known Survival Probability

When the probabilities $\Pr(T_{ij} \leq t|\mathbf{L} = \mathbf{l}, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \boldsymbol{\omega})$ and $\Pr(T_{ij} \leq t|\mathbf{L} = \mathbf{l}, \mathbf{A}_i = \mathbf{a}_i, \boldsymbol{\omega})$ are known beforehand, then the only parameters that need to be estimated are $\mu(t, a, \alpha)$ and $\mu(t, \alpha)$. There are no other nuisance parameters. The following proposition suggests that the proposed estimators are unbiased when the probabilities $\Pr(T_{ij} \leq t|\mathbf{L} = \mathbf{l}, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \boldsymbol{\omega})$ and/or $\Pr(T_{ij} \leq t|\mathbf{L} = \mathbf{l}, \mathbf{A}_i = \mathbf{a}_i, \boldsymbol{\omega})$ are known. The proof is given in Section 3.6

Proposition 2. *If the outcome probabilities* $\Pr(T_{ij} \leq t|\mathbf{L} = \mathbf{l}, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \boldsymbol{\omega})$ *are known, then* $E\{m_i^{int}(t, a, \alpha, \hat{\boldsymbol{\omega}})\} = \mu(t, a, \alpha)$ *and if the outcome probabilities* $\Pr(T_{ij} \leq t|\mathbf{L} = \mathbf{l}, \mathbf{A}_i = \mathbf{a}_i, \boldsymbol{\omega})$ *are known, then* $E\{m_i^{int}(t, \alpha, \hat{\boldsymbol{\omega}})\} = \mu(t, \alpha)$.

Unknown Survival Probability

For unknown outcome probabilities, the distribution of the outcome can be modeled. According to Munda et al. (2012) the conditional hazard $g_{ij}^{OR}(t|\mathbf{L}_{ij}, A_{ij}, \mathbf{A}_{i,-j}, e_i^{OR})$ of $T_{ij}(\mathbf{a}_i)$ can be assumed to be of the following form according to a parametric frailty model

$$g_{ij}^{OR}(t|\mathbf{L}_{ij}, A_{ij}, \mathbf{A}_{i,-j}, \boldsymbol{\omega}, e_i^{OR}) = g_0^{OR}(\boldsymbol{\omega}_h, t)e_i^{OR} \times \exp\left(\mathbf{L}_{ij}^T\boldsymbol{\omega}_{c(1:p_1')} + \phi(\mathbf{A}_{i,-j})\boldsymbol{\omega}_{c(p_1'+1:p')} + A_{ij}\omega_{c(p'+1)}\right).$$

Here g_0^{OR} is the baseline hazard function, $\boldsymbol{\omega}_h$ is the p'' dimensional vector of parameters of the baseline hazard function, e_i^{OR} is a random component following density $f_e^{OR}(e_i^{OR}; \omega_r)$, $\boldsymbol{\omega}_c = (\boldsymbol{\omega}_{c(1:p_1')}, \boldsymbol{\omega}_{c(p_1'+1:p')}, \omega_{c(p'+1)})$ is the vector of coefficients having length $p' + 1$ and ω_r is a parameter of the random effect model. The nuisance parameter vector $\boldsymbol{\omega} = (\boldsymbol{\omega}_c, \boldsymbol{\omega}_h, \omega_r)$ is to be estimated from the model. The function ϕ is a function of the treatment assignment vector possibly included in the model. Under assumption III, the contribution of group i towards the marginal log-likelihood equals (Berg and Drepper

2012)

$$
\begin{aligned}
&l(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\omega}) \\
&\quad = \left(\sum_{j=1}^{n_i} \Delta_{ij} \left[\log\{g_0^{OR}(X_{ij})\} + \mathbf{L}_{ij}^T \boldsymbol{\omega}_{c(1:p_1')} + \phi(\mathbf{A}_{i,-j}) \boldsymbol{\omega}_{c(p_1'+1:p')} + A_{ij}\omega_{c(p'+1)} \right] \right) + \\
&\quad \left[(-1)^{d_i} \mathcal{L}^{(d_i)} \left\{ \sum_{j=1}^{n_i} g_0^{OR}(X_{ij}) \exp\left(\mathbf{L}_{ij}^T \boldsymbol{\omega}_{c(1:p_1')} + \phi(\mathbf{A}_{i,-j}) \boldsymbol{\omega}_{c(p_1'+1:p')} + A_{ij}\omega_{c(p'+1)} \right) \right\} \right]
\end{aligned}
$$

Where $d_i = \sum_{j=1}^{n_i} \Delta_{ij}$ denotes the number of uncensored cases, $\mathcal{L}^{(q)}$ denotes the q-th derivative of the Laplace transform of the frailty distribution, i.e.,

$$
\mathcal{L}^{(s)} = \int_0^{\infty} \exp\left(-e_i^{OR} s\right) f(e_i^{OR}, \omega_r) de_i^{OR}, \qquad s \geq 0.
$$

The estimation of the parameters of the frailty model can be done by maximizing this log-likelihood function. This can also be represented in terms of estimating equations. The maximum likelihood estimate for $\boldsymbol{\omega}$ is a solution of the following estimating equations

$$
\sum_i \psi_{ck}^{OR}(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\omega}) = 0 \text{ for } k = 1, ..., p' + p'' + 1
$$

Here $\psi_{ck}^{OR} = \psi_{ck}^{OR}(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\omega}) = \partial l(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\omega}) / \partial \omega_k$, ω_k being the k-th member of $\boldsymbol{\omega}$. Once these parameters have been estimated, the quantity $\Pr(T \leq t | \mathbf{L}_i, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}})$ required to calculate the proposed estimator can be obtained by integrating out the effect of the random component. If we assume that the estimated survival function corresponding to the model for group i is $S_g(g_{ij}^{OR}(t|\mathbf{L}_i, A_{ij}, \mathbf{A}_{i,-j}, e_i^{OR}, \hat{\boldsymbol{\omega}}))$, then,

$$
\begin{aligned}
&\Pr(T_{ij} \leq t | \mathbf{L}_i, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}}) \\
&\qquad = 1 - \int S_g(g_{ij}^{OR}(t|\mathbf{L}_i, A_{ij} = a, \mathbf{A}_{i,-j}, e_i^{OR}, \hat{\boldsymbol{\omega}})) f_e^{OR}(e_i^{OR}; \omega_r) de_i^{OR}
\end{aligned}
$$

The outcome is assumed to follow a parametric frailty model with a random effect component. Specifically, in the simulation performed and the data analyzed, times to events are assumed to have a gamma frailty distribution. i.e. e_i^{OR} follows gamma distribution with variance ω_r. In this case, a closed form of expression for $\Pr(T_{ij} \leq t|\mathbf{L}_i, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}})$ can be obtained by integration. It can be shown easily by integration that

$$\begin{aligned}
&\Pr(T_{ij} \leq t|\mathbf{L}_i, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}}) \\
&\qquad = 1 - \int \exp\left\{-G_0^{OR}(\mathbf{L}_i, \mathbf{A}_{i,-j}, a, \hat{\boldsymbol{\omega}}, t)e_i^{OR}\right\} \frac{(e_i^{OR})^{1/\hat{\omega_r}-1} e^{-e_i^{OR}/\hat{\omega_r}}}{\hat{\omega_r}^{1/\hat{\omega_r}}\Gamma 1/\hat{\omega_r}} de_i^{OR} \\
&\qquad = 1 - \left\{\frac{1}{\hat{\omega_r} G_0^{OR}(\mathbf{L}_i, \mathbf{A}_{i,-j}, a, \hat{\boldsymbol{\omega}}, t) + 1}\right\}^{1/\hat{\omega_r}}
\end{aligned}$$

Where,

$$\begin{aligned}
G_0^{OR}(\mathbf{L}_i, \mathbf{A}_{i,-j}, a, \hat{\boldsymbol{\omega}}, t) = \exp\left(\mathbf{L}_{ij}^T \hat{\boldsymbol{\omega}}_{c(1:p_1')} + \phi(\mathbf{A}_{i,-j})\hat{\boldsymbol{\omega}}_{c(p_1'+1:p')} + a\hat{\omega}_{c(p'+1)}\right) \\
\times \int_0^t g_0^{OR}(\hat{\boldsymbol{\omega}}_h, s)ds
\end{aligned}$$

Hence, in this case,

$$\begin{aligned}
&m_i^{int}(t, a, \alpha, \hat{\boldsymbol{\omega}}) \\
&\qquad = 1 - \frac{1}{n_i}\sum_{j=1}^{n_i} \sum_{\mathbf{a}_{i,-j} \in \mathcal{A}(n_i-1)} \left\{\frac{1}{\hat{\omega_r} G_0^{OR}(\mathbf{L}_i, \mathbf{A}_{i,-j}, a, \hat{\boldsymbol{\omega}}, t) + 1}\right\}^{1/\hat{\omega_r}} \pi(\mathbf{a}_{i,-k}, \alpha)
\end{aligned}$$

and

$$m^{int}(t, a, \alpha, \hat{\boldsymbol{\omega}}) = \frac{1}{m}\sum_{i=1}^{m} m_i^{int}(t, a, \alpha, \hat{\boldsymbol{\omega}})$$

Equivalently, it can be shown that

$$m_i^{int}(t, \alpha, \hat{\boldsymbol{\omega}}) = 1 - \frac{1}{n_i}\sum_{j=1}^{n_i} \sum_{\mathbf{a}_i \in \mathcal{A}(n_i)} \sum_{a'=0}^{1} \left\{\frac{1}{\hat{\omega_r} G_0^{OR}(\mathbf{L}_i, \mathbf{A}_{i,-j}, a', \hat{\boldsymbol{\omega}}, t) + 1}\right\}^{1/\hat{\omega_r}} \pi(\mathbf{a}_i, \alpha)$$

and

$$m^{int}(t,\alpha,,\hat{\boldsymbol{\omega}}) = \frac{1}{m}\sum_{i=1}^{m} m_i^{int}(t,\alpha,,\hat{\boldsymbol{\omega}})$$

Stacking all the estimating equations together, the following vector of estimating equations are obtained

$$\sum_i \psi^{OR}(\mathbf{O}_i, \boldsymbol{\theta}^{OR}) = 0$$

Here, $\mathbf{O}_i = (\mathbf{X}_i, \mathbf{A}_i, \boldsymbol{\Delta}_i, \mathbf{L}_i)$ is the observed data and $\psi^{OR}(\mathbf{O}_i, \boldsymbol{\omega}) = \left(\boldsymbol{\psi}_{\boldsymbol{c}}^{\boldsymbol{OR}}, \psi_{a\alpha}^{OR}\right)^T$ where $\boldsymbol{\psi}_{\boldsymbol{c}}^{\boldsymbol{OR}} = \left(\psi_{c1}^{OR}, \psi_{c2}^{OR}, ..., \psi_{c(p'+p''+1)}^{OR}\right)^T$, and $\psi_{a\alpha}^{OR} = \psi_{a\alpha}^{OR}(\mathbf{O}_i, \boldsymbol{\omega}, \theta_{a\alpha}^{OR}) = m_i^{int}(t,a,\alpha,\boldsymbol{\omega}) - \theta_{a,\alpha}^{OR}$. The parameters of interest are represented as the vector $\boldsymbol{\theta}^{OR} = \left(\boldsymbol{\omega}, \theta_{a\alpha}^{OR}\right)$. The estimating equations have the vector of solutions $\hat{\boldsymbol{\theta}}^{OR} = \left(\hat{\boldsymbol{\omega}}, \hat{\theta}_{a\alpha}^{OR}\right)$. Denote by $\boldsymbol{\theta}_{\mathbf{0}}^{OR} = \{\boldsymbol{\omega}_0, \mu(t,a,\alpha)\}$, the true value of the parameter $\boldsymbol{\theta}^{OR}$. Then, from proposition 2, $E(m_i^{int}(t,a,\alpha,\boldsymbol{\omega}_0)) = \mu(t,a,\alpha)$ i.e., $E\left(\psi_{a\alpha}^{OR}(\mathbf{O}_i, \boldsymbol{\omega}_0, \mu(t,a,\alpha))\right) = 0$. This can also be expressed as $\int \psi_{a\alpha}^{OR}(\mathbf{o}, \boldsymbol{\omega}_0, \mu(t,a,\alpha))dF_O(\mathbf{o}) = 0$. Along with maximum likelihood theory, this implies the following $\int \psi^{OR}(\mathbf{o}, \boldsymbol{\theta}_0^{OR})dF_O(\mathbf{o}) = 0$.

Using the M-estimation theory put forth by Stefanski and Boos (2002), $\hat{\boldsymbol{\theta}}^{OR} \xrightarrow{p} \boldsymbol{\theta}_{\mathbf{0}}^{OR}$ and $\sqrt{m}(\hat{\boldsymbol{\theta}}^{OR} - \boldsymbol{\theta}_{\mathbf{0}}^{OR})$ converges in distribution to a normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{OR}}$ for a fixed t. Hence, for a fixed time point t, the estimate is consistent and asymptotically normal. The covariance matrix is equal to $U(\boldsymbol{\theta}_0^{OR})^{-1}V(\boldsymbol{\theta}_0^{OR})\{U(\boldsymbol{\theta}_0^{OR})^{-1}\}^T$ where $U(\boldsymbol{\theta}_0^{OR}) = E\{-\dot{\psi}^{OR}(\mathbf{O}_i, \boldsymbol{\theta}_0^{OR})\}$, $V(\boldsymbol{\theta}_0^{OR}) = E\{\psi^{OR}(\mathbf{O}_i, \boldsymbol{\theta}_0^{OR})\psi^{OR}(\mathbf{O}_i, \boldsymbol{\theta}_0^{OR})^T\}$, and $\dot{\psi}^{OR}(\mathbf{O}_i, \boldsymbol{\theta}^{OR}) = \partial\psi^{OR}(\mathbf{O}_i, \boldsymbol{\theta}^{OR})/\partial(\boldsymbol{\theta}^{OR})^T$. So $m(t,a,\alpha,\hat{\boldsymbol{\omega}})$ is consistent and asymptotically normal. Similar techniques can be used to show that $m(t,\alpha,\hat{\boldsymbol{\omega}})$ is also consistent and asymptotically normal. The asymptotic variance $\boldsymbol{\Sigma}^{OR}$ can be consistently estimated by $\hat{\boldsymbol{\Sigma}}^{OR} = \hat{U}(\hat{\boldsymbol{\theta}}^{OR})^{-1}\hat{V}(\hat{\boldsymbol{\theta}}^{OR})\{\hat{U}(\hat{\boldsymbol{\theta}}^{OR})^{-1}\}^T$ where $\hat{U}(\hat{\boldsymbol{\theta}}^{OR}) = m^{-1}\sum_{i=1}^{m}\{-\dot{\psi}^{OR}(\mathbf{O}_i, \hat{\boldsymbol{\theta}}^{OR})\}$ and $\hat{V}(\hat{\boldsymbol{\theta}}^{OR}) = m^{-1}\sum_{i=1}^{m}\{\psi^{OR}(\mathbf{O}_i, \hat{\boldsymbol{\theta}}^{OR})\psi^{OR}(\mathbf{O}_i, \hat{\boldsymbol{\theta}}^{OR})^T\}$. The sandwich variance estimator $\hat{\boldsymbol{\Sigma}}^{OR}$ can be computed using the R package `geex` (Saul and Hudgens 2017) and can be used to construct Wald type confidence intervals.

3.3 Simulation

To demonstrate the efficacy of the methods discussed, a simulation study was performed. The data were simulated using the following steps

1. First, baseline covariate L_{1ij} was generated randomly. Random variables V_{ij} were generated following an exponential distribution with mean 20. The baseline covariates L_{1ij} were then assigned to be $\min\{V_{ij}, 100\}/10$.

2. Random effects b_i for generating the treatment probabilities were randomly generated following a normal distribution with mean 0 and variance 1.0859.

3. Treatment indicators were simulated following a Bernoulli distribution with probability p_{ij} where $p_{ij} = \text{expit}(0.2727 - 0.0387L_{1ij} + b_i)$.

4. Random components b'_i for generating the potential times to event were generated randomly from a gamma distribution with mean 1 and variance 10.

5. Then the time to events $T_{ij}(\mathbf{a}_i)$ were randomly sampled from an exponential distribution with mean μ_{ij} where $1/\mu_{ij} = b'_i \exp\left(-3.1a_{ij} - 0.2\sum_{k\neq j} a_{ik} + 5.3L_{1ij}\right)$.

6. Next, another set of random effects e_i^{OR} were generated randomly following a gamma distribution with mean 1 and variance 0.1

7. Finally, the censoring times C_{ij} were sampled randomly from an exponential distribution with mean λ_0 where where $\lambda_0 = 0.1\exp\left(0.01L_{1ij}\right)e_i^{OR}$.

8. An individual was censored, i.e. $\Delta_{ij} = 1$ if $C_{ij} < T_{ij}(\mathbf{A}_i)$.

Data were simulated from 250 groups with each group having 30 individuals using the steps of the simulation. Hence each of the simulated data set had 7500 individuals in the sample. The steps of the simulation were performed 10,000 times. To check the indirect effects, several different values of the treatment allocation probability α were

used for estimation. Each of the 10,000 simulations generated estimates of the quantities of interest. These estimates were compared with the true value of the estimates obtained using counterfactual data generated for a large data set. It can be observed that since the allocation probabilities are Bernoulli, whichever values of the vector $\mathbf{a}_{i,-j}$ yield the same value of $\sum_{k\neq j} a_{ik}$, also produce the same value for the quantity $\pi(\mathbf{a}_{i,-j}, \alpha)$. So, according to the data generation mechanism, the population average potential cumulative survival distribution at the time point t for treatment a under allocation strategy α can be represented as follows (Perez-Heydrich et al. 2014)

$$m^{-1}\sum_i n_i^{-1}\sum_j\sum_{k=0}^{n_i-1}\binom{n_i-1}{k}I\{T_{ij}(a,\mathbf{a}_{i,-j})\leq t\}\alpha^k(1-\alpha)^{n_i-k-1}$$

Similarly, the population average marginal potential cumulative survival distribution at the time point t under allocation strategy α can be represented as follows

$$m^{-1}\sum_i n_i^{-1}\sum_j\sum_{k=0}^{n_i}\binom{n_i}{k}I\{T_{ij}(a,\mathbf{a}_{i,-j})\leq t\}\alpha^k(1-\alpha)^{n_i-k-1}$$

This representation is possible since the data generation mechanism ensures that the potential time to events are only dependent on the vector $\mathbf{a}_{i,-j}$ through the sum of its elements. A sandwich variance estimator of the proposed estimators was also obtained for each of the simulated data. For each value of α, the estimated values of the quantities of interest are calculated by averaging over all the simulated data set. Asymptotic standard errors were also calculated by taking the mean of the sandwich variance estimator. Empirical standard errors were calculated as the standard deviation of the estimates. 95% Wald confidence intervals were constructed and coverage probabilities for each value of α were calculated by taking the proportion of estimators with values within the confidence interval. Results of the simulation are summarized in Table 3.1, which describes the results obtained in detail for nine different α values at time point 100 days.

3.4 Data

3.4.1 Cholera Vaccine Study and Analysis

The methods described in Section 3.2 are applied to a cholera vaccine study in Matlab from a cholera vaccine trial in Matlab, Bangladesh (Ali et al. 2005). Children of ages 2-15 and women were the participants in the study. The vaccine administered was either B subunit-killed whole-cell oral cholera vaccine or killed whole-cell-only cholera vaccine and the placebo administered was E. coli K12 placebo. For analysis, the two versions of the treatments were assumed to be the same and the study participants were randomized to receive one of the three treatments with equal probability. An individual was considered a participant only if he/she received two or more doses of the treatment assigned to him/her. A participation vector was used to keep track of the participants and non-participants. The doses of treatment and vaccines were administered during the months of January to May in the year 1985. Three centers were established for vaccination purposes and maintained as surveillance centers. The data consisted of a total of 121,982 individuals. Previous studies have managed to establish that interference is present within the data (Perez-Heydrich et al. 2014). But, the issue of censoring in the data has not been addressed in any of these studies to date.

The data is right censored because the time to having an event of cholera is not observed for all the individuals. Censoring could be due to an unobserved event within the duration of the study, migration from study location or death during the follow up period.

The data were readied for analysis using the following steps. 7 of the 121,982 observed individuals appeared to be duplicate values and consequently were removed from the data giving rise to a data with $121,975$ individuals. As mentioned before, the vaccines were administered during a five-month window. So, even if the date of entry of individuals in the study were not exactly the same, they were comparable. So, the same date of entry

was assigned to all the individuals within a particular group for ease of computation. The representative start date for a group was assumed to be the most recent date of second vaccination among all the individuals within that group. In case all of the individuals failed to receive a second vaccination, the representative start date for a group was then assumed to be the most recent date of first vaccination among all the individuals within that group. After assigning start dates to all the individuals, it was observed that 34 individuals could not be assigned any start date. Deleting those 34 data observations 121,941 individuals remained within the data. Among them, the representative start dates for 60 were after the date of contracting cholera which meant that they had a negative time to event. So those observations could not be used and were deleted from the data. Again, from the remaining data, 3617 were lost due to migration and 346 were lost due to death before the representative start date of their group. The remaining 117,918 individuals constituted the final dataset used for analysis. Placebo was administered to a total of 69,219 individuals and treatment to 48,699 individuals within the final data. Only 375 were cases of cholera recorded in the control group and 103 in the treatment group.

As discussed in Section 3.2.3, for fitting the parametric frailty model to the outcome of interest, i.e., time to cholera, both the baseline hazard function and the frailty distribution must be specified. There are several choices for the baseline hazard and frailty model combination.

Table 3.2 summarizes AIC and BIC values for several baseline hazard function and frailty distribution combinations. From Table 3.2, it can be observed that both AIC and BIC are minimum for gompertz baseline hazard with gamma frailty distribution. Therefore, in order to fit the outcome model to the Matlab cholera data, a parametric frailty model with baseline hazard distribution gompertz and frailty distribution gamma was selected. For gompertz baseline hazard, the function $g_0^{OR}(\boldsymbol{\omega}_h, t)$ equals $\omega_{h2} exp(\omega_{h1} t)$. The

covariates included in the outcome model were treatment status, proportion of individuals treated in the group, interaction of treatment status and proportion of individuals treated within the group, age, distance to river, squared age and squared distance to river. Also, as shown in Section 3.2.3, both the outcome probability as well as the likelihood function has a closed form expression in this case. As more than 70% of the individuals in the data were in groups where the proportion of people treated were in between 0.3 and 0.6, the analysis was restricted for those values of α.

3.4.2 Results

Figure 3.1 shows the estimated cumulative probability distribution obtained using the parametric g formula compared against the IPCW method over time for three different α values. Although the parametric g formula and the IPCW method do not produce exactly the same curves, the overall trends of both are quite similar. The jumps in probability for the IPCW estimators are more pronounced than the parametric g formula estimators after some time. According to the parametric g formula, the cumulative probability seems to increase at an exponential rate for both the vaccinated and unvaccinated groups. But the rate of increase is much larger for the control group compared to the vaccine group. Therefore, the difference in the cumulative probability of cholera between the control and vaccine group seems to increase with time. So, it can be deduced that the vaccine effects get more and more pronounced with time. The direct effect of vaccination becomes increasingly significant over time. However, changes in the policy α also seem to affect the cumulative probability of cholera. For a particular time point, it can be observed from Figure 3.1 that, the cumulative probability of cholera decreases with the increment of allocation strategy α. This reduction is more prominent in the control group than the vaccine group. This hints at the presence of an indirect treatment effect which affects untreated individuals as well. This indirect effect or spillover effect or herd immunity can be explained through interference.

Figure 3.2 depicts the direct, indirect, total and overall effects at one year after the start of the study multiplied by 1000. Among all the effects, only the direct effect has a decreasing trend with the allocation probability. The rest of the effects show an increasing trend. The direct effect is greater than 0 for all the policy α in between 0.3 and 0.6. For example, the estimates of the direct effect of treatment at allocation level $\alpha = 0.4$ is 2.2 with 95% confidence interval $(1.5, 2.9)$ and the direct effect of treatment at allocation level $\alpha = 0.5$ is 1.5 with 95% confidence interval $(0.9, 2.1)$. None of the confidence intervals contain 0 and the higher value of α gives rise to lesser direct effect. Also, it can be said that for policy $\alpha = 0.4$, the expected number of people contracting cholera is 2.2 more in the unvaccinated group compared to the vaccinated group per 1000 individual. The indirect, total and overall effects are measures of the spillover effect of the vaccine. From Figure 3.2 it is apparent that higher allocation probabilities translate to greater spillover effects. For allocation probabilities less than 0.4, the indirect and overall effects are less than 0 and they are greater than 0 for allocation probabilities greater than 0.4. As an example, the estimate of the overall effect corresponding to allocation levels $\alpha_1 = 0.4$ and $\alpha_2 = 0.6$ is 1.5 with confidence 95% confidence interval $(1.0, 2.1)$. So, 1.5 more individuals are expected to contract cholera if they belong to a neighborhood with policy $\alpha = 0.4$ compared to a neighborhood with policy $\alpha = 0.6$ per 1000 individuals.

The Matlab cholera vaccine data has been analyzed previously and causal effects were calculated at one year of follow up. (Perez-Heydrich et al. 2014) used an inverse probability weighted method using group weights instead of individual weights. But they failed to account for the right censoring present within the data. However, the effect plots show a somewhat similar trend. As an example, for policy level $\alpha = 0.32$, estimate of the direct effect according to Perez-Heydrich et al. was calculated to be 5.3 and the 95% confidence interval was given to be $(2.5, 8.1)$. The parametric g estimate of the direct effect at allocation level $\alpha = 0.32$ as shown in Figure 3.2 is 2.8 having 95% confidence interval $(1.7, 3.9)$. Even though the confidence interval is much wider for the estimate

provided by Perez-Heydrich et al., still none of the confidence intervals include 0 and both the effects are greater than 0 again signifying significant direct effect of the vaccine.

3.5 Discussion

In this section, a new method of calculating causal effect for data with interference and right censoring has been proposed. The method involves fitting an outcome regression model and using standardization to get a parametric g formula estimate for various causal effects. The outcome model fitted was a parametric frailty model and the outcome probability was calculated by integrating out the random effect. The causal effects discussed in this section are direct effects, indirect effects, total effects and overall effects. It was proved that the parametric g formula estimator is consistent and asymptotically normal. Also, the variance matrix was estimated consistently using a sandwich variance estimator. 95% confidence intervals were also constructed for the point estimates using these sandwich variance estimators. A simulation study was performed and results were compared to the true values of the parameters to observe the finite sample performance of the methods. Finally, causal effect estimates and 95% confidence intervals were calculated using the parametric g formula methods for a cholera vaccine study conducted in Matlab, Bangladesh.

The outcome model in this section is specified to be a parametric frailty model. However, it might be possible to extend these methods and use a Cox proportional hazards model with a random component instead. Also, a possible future direction of research might be to try an extend the methods discussed in this section for general interference instead of partial interference. For analyzing the real data, the outcome model was selected based on AIC and BIC. It might be possible to come up with a better method of model selection. The calculation of parametric g formula involves computing the sum $\sum_{\mathbf{a}_i \in \mathcal{A}(n_i)} \Pr(T \leq t | \mathbf{L}_i, \mathbf{A}_i = \mathbf{a}_i, \hat{\boldsymbol{\omega}})$. Summing over all possible values of $\mathbf{a}_i$ might be computationally difficult. In that case, a Monte-Carlo approach used by Liu et al.

(2018) might be used to relieve the computational burden. Another shortcoming of this method is that the outcome model is sensitive to model misspecification. The estimates can be misleading if the model is not specified correctly. Therefore, one possible direction for future work might be to explore doubly robust estimators which might be stable under model misspecification under some regularity conditions.

Figure 3.1: Estimated cumulative probability of cholera against time for vaccine and control for $\alpha = 0.3$ (left), $\alpha = 0.45$ (center) and $\alpha = 0.6$ (right)

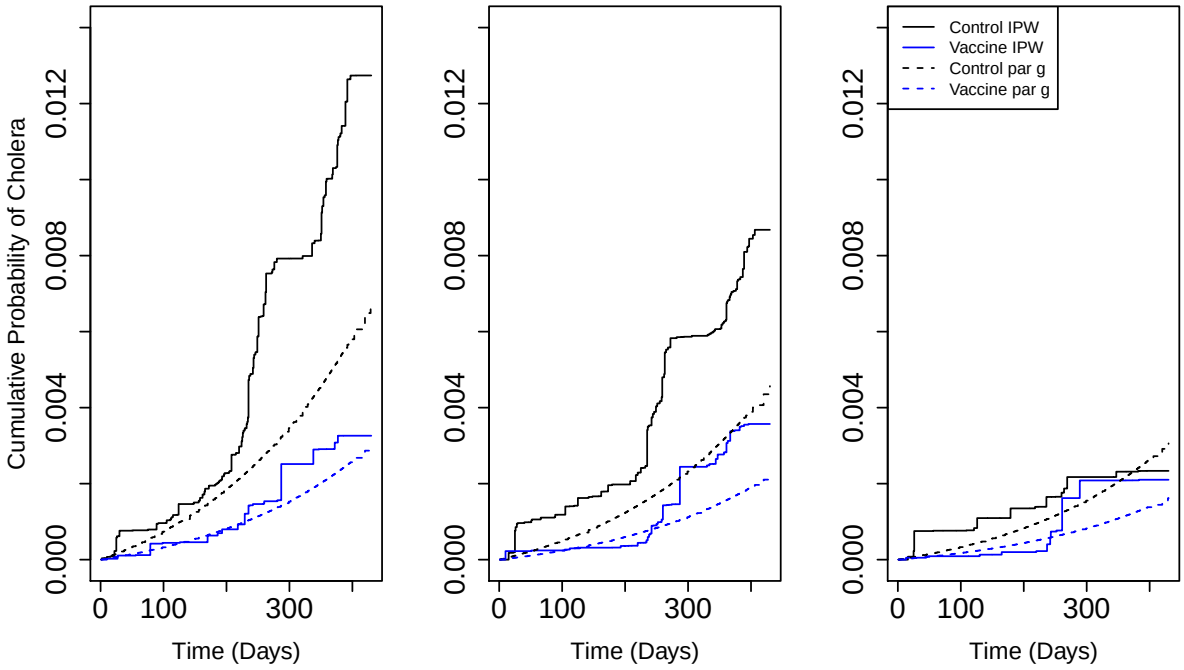
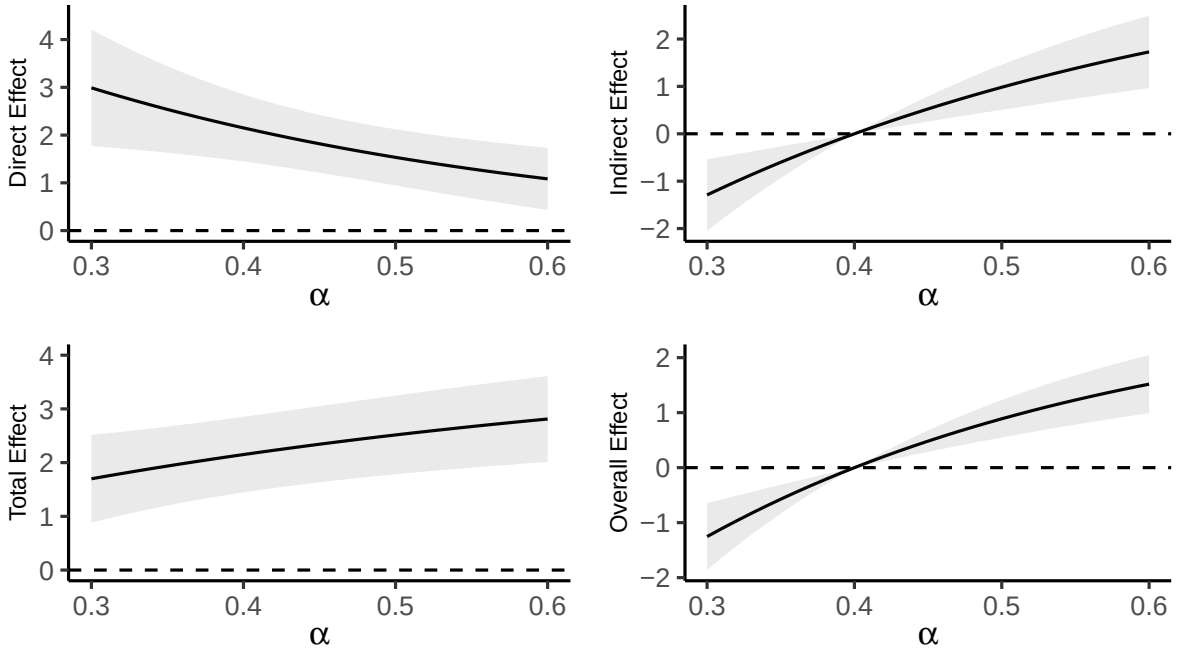


Figure 3.2: Direct effect, indirect effect, total effect and overall effect estimates multiplied by 1000 for different allocation strategies at time $t = 1$ year. Indirect effects, total effects and overall effects are with respect to $\alpha_2 = 0.4$. The shaded region denotes the 95% confidence interval of the estimates.



α	$\mu(100,0,\alpha)$	Bias	ESE	ASE	EC	α	$\mu(100,1,\alpha)$	Bias	ESE	ASE	EC
0.1	0.348	-0.000	0.025	0.023	94%	0.1	0.469	-0.000	0.025	0.023	94%
0.2	0.368	-0.000	0.023	0.022	94%	0.2	0.494	-0.000	0.022	0.020	94%
0.3	0.391	-0.000	0.022	0.020	94%	0.3	0.519	-0.000	0.019	0.018	94%
0.4	0.413	-0.000	0.020	0.019	95%	0.4	0.543	0.000	0.017	0.016	95%
0.5	0.437	-0.000	0.019	0.018	95%	0.5	0.566	-0.000	0.016	0.015	95%
0.6	0.461	-0.000	0.018	0.018	95%	0.6	0.589	-0.000	0.015	0.014	95%
0.7	0.486	-0.000	0.019	0.018	95%	0.7	0.611	-0.000	0.015	0.014	95%
0.8	0.510	-0.000	0.019	0.018	94%	0.8	0.631	-0.000	0.016	0.015	94%
0.9	0.535	-0.000	0.021	0.020	94%	0.9	0.651	-0.000	0.017	0.016	94%

Table 3.1: Results from simulation study described in Section 3.3. α denotes the allocation probabilities, $\mu(100, a, \alpha)$ is the true value of the survival probabilities at time point 100 for $a = 0, 1$; bias is the average of $\mu(100, a, \alpha) - m^{int}(100, a, \alpha, \hat{\boldsymbol{\omega}})$, ESE is the empirical standard error, ASE is the average of the sandwich variance estimators and EC denotes the empirical coverage of the 95% Wald type confidence intervals.

	Gamma	Inverse Gaussian	Positive Stable
Exponential	5946.42 (6033.51)	5955.73 (6042.83)	5992.52 (6079.61)
Weibull	5888.98 (5985.76)	5898.34 (5995.12)	5935.18 (6031.96)
Gompertz	5858.04 (5954.82)	5867.50 (5964.28)	5904.57 (6001.35)
Loglogistic	5889.35 (5986.13)	5898.70 (5995.48)	5935.41 (6032.19)
Lognormal	7296.60 (7393.38)	7641.41 (7738.18)	6559.83 (6656.61)

Table 3.2: AIC (BIC) values for different baseline hazard functions corresponding to gamma, inverse Gaussian and Positive stable frailty distributions.

CHAPTER 4: DOUBLY ROBUST ESTIMATION FOR DATA WITH PARTIAL INTERFERENCE AND RIGHT CENSORING

4.1 Introduction

Interference is present if the treatment status of an individual affects the outcome of another individual in a data (Cox 1958). Interference can often be observed in data pertaining to infectious diseases. That is because more often than not, the disease status of an individual depends not only on the vaccination status of that particular individual but also on the vaccination status of other individuals (Halloran and Struchiner 1991). When the data can be partitioned into groups such that the members of a group can interfere within themselves but there is no interference between members of any two different groups then this is a sub-case of interference termed as partial interference Sobel (2006). The groups can be based on social, temporal and/or geographical similarities which are apparent from the data. The effect that one might be concerned with due to a data having interference is termed as peer effect or spillover effect. Examples of areas concerned with these effects are criminology (Sampson 2010, Verbitsky-Savitz and Raudenbush 2012), developmental psychology (Duncan et al. 2005, Foster 2010), econometrics (Sobel 2006, Manski 2013) education (Hong and Raudenbush 2006, Vanderweele et al. 2013), imaging (Luo et al. 2012), political science (Sinclair et al. 2012, Bowers et al. 2013), social media and network analysis (VanderWeele and An 2013, Toulis and Kao 2013, Eckles et al. 2014, Kramer et al. 2014), sociology (Aronow and Samii 2017), and spatial analyses (Zigler et al. 2012, Graham et al. 2013).

Various methods of inference has been suggested for randomized experiments under the partial interference setting (e.g., Rosenbaum 2007, Hudgens and Halloran 2008, Baird

et al 2014, Eckles et al 2016). However, randomized experiments might not always be feasible to construct or there might be ethical issues arising in performing a randomized trial. In that case, for observational data, methods proposed by Tchetgen Tchetgen and VanderWeele (TV) has been proven to be useful. TV provided consistent estimators for various causal effects of concern using inverse probability weighting (IPW). But, IPW methods are known to suffer from some serious drawbacks. When the propensity scores required to calculate the inverse probability weights become small, the estimator is highly unstable and difficult to compute numerically. The parametric g formula bypasses this issue and might be used in place of the IPW estimators. Robins (1986) first suggested the use of g-computation algorithm. Using outcome regression with g-computation gives rise to the parametric g formula which is calculated using standardization techniques (Hernán and Robins 2006).

However, both the IPW method and parametric g formula are based on a strong assumption. The assumption is that the model used in each case is specified correctly. Otherwise, the estimators will not be consistent. For the IPW estimator, the treatment model must be correct and for the parametric g formula, the outcome model must be specified correctly. The doubly robust estimator tries to address this issue by incorporating robustness. For these estimators, only one of the treatment and outcome model must be specified correctly for getting consistent estimators. In that sense, the estimator is robust under model misspecification.

The rest of this section is organized as follows- Section 4.2 discusses the doubly robust method in detail and develops large sample properties of the estimator, Section 4.3 provides results for various simulation scenarios implemented, data from a cholera vaccine study conducted in Matlab, Bangladesh is analyzed in Section 4.4, and finally, in Section 4.5, a brief discussion of the whole section is provided.

4.2 Methods

4.2.1 Estimands

Suppose that the data consists of m groups with group i having n_i individuals for $i = 1, \ldots, m$. The treatment status indicator A_{ij} equals 1 when individual j in group i receives treatment and $A_{ij} = 0$ if said individual receives placebo. Represent the vector $(A_{i1}, A_{i2}, ..., A_{in_i})$ by $\mathbf{A}_i$ and the vector $(A_{i1}, A_{i2}, ..., A_{ij-1}, A_{ij+1}..., A_{in_i})$ by $\mathbf{A}_{i,-j}$. The random variables $\mathbf{A}_i$ and $\mathbf{A}_{i,-j}$ can take values $\mathbf{a}_i$ and $\mathbf{a}_{i,-j}$ respectively. The notation $T_{ij}(\mathbf{a}_i)$ represents the potential time to event for individual j in group i with treatment $\mathbf{a}_i$. Let $\mathbf{T}_i(.) = \{T_{ij}(\mathbf{a}_i) : \mathbf{a}_i \in \mathcal{A}(n_i), j = 1, 2, \cdots, n_i\}$ denote the set of all potential event times for individuals in group i. Also assume that because of loss to follow up, completion of study or other reasons, the time to events might be right censored for some observations. Then the censoring times for individual j in group i might be denoted by C_{ij}. Further assume that the indicator of censoring $\Delta_{ij} = I(T_{ij}(\mathbf{A}_i) \leq C_{ij})$ and the observed time to events $X_{ij} = \min(T_{ij}(\mathbf{A}_i), C_{ij})$. Denote $\mathbf{X}_i = (X_{i1}, X_{i2}, \cdots, X_{in_i})$ and $\mathbf{\Delta}_i = (\Delta_{i1}, \Delta_{i2}, \cdots, \Delta_{in_i})$. The set containing all 2^{n_i} combinations of treatments in a group with n_i observations, is termed as $\mathcal{A}(n_i)$ for $n_i = 1, 2, \ldots$. Consider the vector of possible covariates to be $\mathbf{L}_{ij}$ for individual j in group i. Assume the group sizes n_i to be random variables included in the vector of baseline covariate $\mathbf{L}_{ij}$. For a particular group i, the matrix of all possible covariates combining all the covariates of the individuals in the group i is denoted by $\mathbf{L}_i$. The overall observed data can be viewed as m i.i.d observations of $(\mathbf{L}_i, \mathbf{A}_i, \mathbf{X}_i, \mathbf{\Delta}_i)$ for $i = 1, 2, \ldots, m$.

There are various causal effects that are of interest. These causal effects are often represented as different contrasts of expected potential outcomes. When interference is absent, a popular causal effect of interest is the average causal effect which is defined as the difference of the expected potential outcomes for treated and untreated individuals.

In case of interference, various causal effects can be defined along similar lines by considering contrasts under various counterfactual outcomes (Hong and Raudenbush 2006, Sobel 2006, Hudgens and Halloran 2008, Tchetgen and VanderWeele 2012). An additional term of concern for data with interference is the spillover effect or herd immunity which is not taken into consideration for data without interference. These gives rise to various casual effects which are not of concern in absence of interference. To define these additional causal effects, first, the allocation probability α needs to be introduced. Allocation probability or policy α is defined as the probability of receiving treatment for an individual independent of the others. TV interpreted this term as the independent Bernoulli probability α for individuals being assigned treatment. So, under ithis counterfactual scenario, everyone receives treatment independently f others with probability α. According to the Bernoulli probability assignment, the probability of a group treatment vector can be calculated using independence of treatment assignment. So, denoting by $\pi(\mathbf{a}_i, \alpha)$, the probability that group i has treatment vector $\mathbf{a}_i$ under allocation policy α, it can be deduced that $\pi(\mathbf{a}_i, \alpha) = \Pr_\alpha(\mathbf{A}_i = \mathbf{a}_i) = \prod_{k=1}^{n_i} \alpha^{a_{ik}}(1-\alpha)^{1-a_{ik}}$. Similarly, denoting by $\pi(\mathbf{a}_{i,-j}, \alpha)$, the probability that group i except for individual j has treatment vector $\mathbf{a}_{i,-j}$, it can be said that $\pi(\mathbf{a}_{i,-j}, \alpha) = \Pr_\alpha(\mathbf{A}_{i,-j} = \mathbf{a}_{i,-j} | A_{i,j} = a) = \prod_{k=1,k\neq j}^{n_i} \alpha^{a_{ik}}(1-\alpha)^{1-a_{ik}}$. Here $\Pr_\alpha(\cdot)$ corresponds to the probability under this counterfactual setting.

The causal effects of interest at a particular time point t are defined as contrasts of population average potential cumulative probability or population average marginal potential cumulative probability of an event before time t. To define these, first the average probability of individual j in group i observing an event by time t when said individual receives treatment a and the allocation strategy of group i is α has to be defined as follows

$$\bar{F}_{ij}(t, a, \alpha) = \sum_{a_{i,-j} \in \mathcal{A}(n_i - 1)} I\{T_{ij}(a, \mathbf{a}_{i,-j}) \leq t\} \pi(\mathbf{a}_{i,-j}, \alpha).$$

Similarly, the marginal average probability of individual j observing an event in group i by time t when said individual receives treatment a and the allocation strategy of group i is α can be defined as

$$\bar{F}_{ij}(t,\alpha) = \sum_{a_i \in \mathcal{A}(n_i)} I\{T_{ij}(\mathbf{a}_i) \leq t\}\pi(\mathbf{a}_i,\alpha).$$

Then, let $\bar{F}_i(t,a,\alpha) = n_i^{-1}\sum_{j=1}^{n_i}\bar{F}_{ij}(t,a,\alpha)$ and $\bar{F}_i(t,\alpha) = n_i^{-1}\sum_{j=1}^{n_i}\bar{F}_{ij}(t,\alpha)$. The quantity $\bar{F}_i(t,a,\alpha)$ is the representation of the average probability of an individual observing an event in group i by time t when said individual receives a and the allocation strategy of group i is α. Similarly, $\bar{F}_i(t,\alpha)$ is the representation of the marginal average probability of an individual observing an event in group i by time t when the allocation strategy of group i is α. Hence, $\mu(t,a,\alpha) = E\{\bar{F}_i(t,a,\alpha)\}$ can be termed as the population average potential cumulative distribution at time point t for treatment a under allocation strategy α. Similarly, $\mu(t,\alpha) = E\{\bar{F}_i(t,\alpha)\}$ can be termed as the population average marginal potential cumulative distribution at time point t under allocation strategy α. The term $\mu(t,a,\alpha)$ has the interpretation as the probability of an individual's time to event being less than t under the counterfactual scenario that an individual receives treatment a with the group allocation probability being α. For example, in the Matlab cholera vaccine study described in section 4.4, $\mu(t,a,\alpha)$ equals the population average potential cumulative probability of an individual to get infected with cholera before time t when the individual receives treatment a and the group treatment allocation probability is α.

Finally, the causal effects of interests can be defined in terms of contrasts of $\mu(t,a,\alpha)$ and $\mu(t,\alpha)$. The primary effects of treatment on the outcome is measured through direct effect which is defined as the contrast between the population average potential cumulative probability for treatment and control for a particular time point and policy. Mathematically, the direct effect of treatment on the outcome at time point t for allocation level α can be given by $DE(t,\alpha) = \mu(t,0,\alpha) - \mu(t,1,\alpha)$. The rest of the

causal effects are concerned with the spillover effect or herd immunity in addition to the direct effect of treatment. For example, the indirect effect is defined as the contrast between the population average potential cumulative probability for two different allocation levels under no treatment for a particular time point. This effect measures how treatment indirectly affects the outcome through herd immunity in the unvaccinated group and is given by $IE(t, \alpha_1, \alpha_2) = \mu(t, 0, \alpha_1) - \mu(t, 0, \alpha_2)$ for allocation strategies α_1 and α_2. The total effect is defined as the contrast between the population average potential cumulative probability for two different allocation levels under treatment and no treatment for a particular time point. This effect can be interpreted as an aggregate measure of the direct and indirect effects of treatment. Mathematically, the total effect of treatment on the outcome at time point t for allocation level α_1 and α_2 can be given by $TE(t, \alpha_1, \alpha_2) = \mu(t, 0, \alpha_1) - \mu(t, 1, \alpha_2)$. So, it can be seen that $TE(t, \alpha_1, \alpha_2) = DE(t, \alpha_2) + IE(t, \alpha_1, \alpha_2)$. Lastly, the overall effect is used to measure the overall effect of the allocation strategy irrespective of the treatment status. Defined as the contrast between the population average marginal potential cumulative probability for two different allocation levels at a particular time point, the overall effect can be viewed as a combination of direct and indirect effects. The overall effect can be written as $OE(t, \alpha_1, \alpha_2) = \mu(t, \alpha_1) - \mu(t, \alpha_2)$. It can be seen that since $\mu(t, \alpha) = (1 - \alpha)\mu(t, 0, \alpha) + \alpha\mu(t, 1, \alpha)$, $OE(t, \alpha_1, \alpha_2) = IE(t, \alpha_1) - \alpha_1 DE(t, \alpha_1) + \alpha_2 DE(t, \alpha_2)$.

4.2.2 Assumptions

Assume the following,

I) Conditional independence: $\mathbf{A}_i \perp\!\!\!\perp \mathbf{T}_i(.)|\mathbf{L}_i$,

II) Positivity: $\Pr(\mathbf{A}_i = \mathbf{a}_i|\mathbf{L}_i = \mathbf{l}) > 0$ for all $\mathbf{a}_i \in \mathcal{A}(n_i)$ and $\mathbf{l}$ such that $\Pr(\mathbf{L}_i = \mathbf{l}) > 0$,

III) Conditional independent censoring: $\mathbf{C}_i \perp\!\!\!\perp \{\mathbf{T}_i(.), \mathbf{A}_i\}|\mathbf{L}_i$.

Assumptions I and II are generalizations of individual level assumptions for interference. Both of these are assumed in case of no interference as well. However, for no interference the assumptions are made on individuals whereas with interference the assumptions must be extended for groups instead of individuals. For example, for no interference, assumption I, commonly referred to as no unmeasured confounding states that the potential outcome for each individual under each particular treatment is independent of the actual treatment assignment of the individual given all of the measured confounders for the individual. A simple extension states that the potential outcome for all individuals of a group is independent of the actual treatment assignment of the group given all of the measured confounders for the group. Similarly, assumption II, termed as positivity, signifies that each group has a positive probability of being assigned every possible treatment combination given all of the measured confounders for the group. Assumption III pertains to the censoring distributions. It states that for each individual, given the set of measured confounders for the individual's group, the potential outcome of a group and the observed group treatment assignment are jointly conditionally independent of the censoring time of for that group given covariates. It can be shown that this assumption can be relaxed further to make the censoring times depend on the treatment status.

4.2.3 IPCW and Parametric G Formula Estimators

For interference, the TV IPW estimators with group propensity scores can be extended with censoring weights to formulate IPCW estimators with group propensity scores to estimate $\mu(t, a, \alpha)$ and $\mu(t, \alpha)$ as follows- $\hat{\mu}(t, a, \alpha) = m^{-1} \sum_{i=1}^{m} \hat{F}_i(t, a, \alpha)$ and $\hat{\mu}(t, \alpha) = m^{-1} \sum_{i=1}^{m} \hat{F}_i(t, \alpha)$ where

$$\hat{F}_i(t, a, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \frac{\pi(\mathbf{A}_{i,-j}; \alpha) I(A_{ij} = a) I(\Delta_{ij} = 1) I(X_{ij} \leq t)}{\Pr(\mathbf{A}_i | \mathbf{L}_i, \hat{\boldsymbol{\beta}}) \Pr(\Delta_{ij} = 1 | \mathbf{L}_i, X_{ij}, \hat{\boldsymbol{\gamma}})},$$

and

$$\hat{F}_i(t,\alpha) = n_i^{-1} \sum_{j=1}^{n_i} \frac{\pi(\mathbf{A}_i;\alpha) I(\Delta_{ij}=1) I(X_{ij} \leq t)}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\hat{\boldsymbol{\beta}}) \Pr(\Delta_{ij}=1|\mathbf{L}_i, X_{ij}, \hat{\boldsymbol{\gamma}})}.$$

The estimated parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are used to calculate the group propensity scores $\Pr(\mathbf{A}_i|\mathbf{L}_i,\hat{\boldsymbol{\beta}})$ and censoring weights $\Pr(\Delta_{ij}=1|\mathbf{L}_i, X_{ij}, \hat{\boldsymbol{\gamma}})$ respectively. Methods for calculating the propensity score model and censoring model are discussed later. From these IPCW estimates ,the various causal effects can also be estimated as follows- $\widehat{DE}(t,\alpha) = \hat{\mu}(t,0,\alpha) - \hat{\mu}(t,1,\alpha)$, $\widehat{IE}(t,\alpha_1,\alpha_2) = \hat{\mu}(t,0,\alpha_1) - \hat{\mu}(t,0,\alpha_2)$, $\widehat{TE}(t,\alpha_1,\alpha_2) = \hat{\mu}(t,0,\alpha_1) - \hat{\mu}(t,1,\alpha_2)$ and $\widehat{OE}(t,\alpha_1,\alpha_2) = \hat{\mu}(t,\alpha_1) - \hat{\mu}(t,\alpha_2)$. Under the assumptions discussed before and when the treatment and censoring models are correctly specified, the IPCW estimator can be shown to be consistent and asymptotically normal.

The group propensity scores $\Pr(\mathbf{A}_i|\mathbf{L}_i,\hat{\boldsymbol{\beta}})$ required to calculate the IPCW estimators might become too small for large groups and numerical calculation might become unfeasible. One way to circumvent this problem might be to use parametric g formula to find estimators for $\mu(t,a,\alpha)$ and $\mu(t,\alpha)$. The parametric g formula estimator for $\mu(t,a,\alpha)$, in this case, can be given by

$$m^{int}(t,a,\alpha,\hat{\boldsymbol{\omega}}) = \int \sum_{\mathbf{a}_{i,-k} \in \mathcal{A}(n_i-1)} \Pr(T \leq t|\mathbf{L}=\mathbf{l}, A=a, \mathbf{A}_{i,-k}=\mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}}) dF_{\mathbf{L}}(\mathbf{l}) \times \pi(\mathbf{a}_{i,-k},\alpha).$$

The outcome probability $\Pr(T \leq t|\mathbf{L}=\mathbf{l}, A=a, \mathbf{A}_{i,-k}=\mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}})$ can be estimated by fitting a parametric model which is discussed in detail in the following sections. Here $\hat{\boldsymbol{\omega}}$ is an estimator of the parameter for the outcome model. The quantity $\Pr(T \leq t|\mathbf{L}=\mathbf{l}, A=a, \mathbf{A}_{i,-k}=\mathbf{a}_{i,-k}, \boldsymbol{\omega})$ is well defined because of assumption II. The integral in the estimator can be replaced by sum and the parametric g formula estimator can be

empirically calculated based on the empirical distribution of $\mathbf{L}$ as-

$$\begin{aligned} m^{int}(t,a,\alpha,\hat{\boldsymbol{\omega}}) &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_{i,-j}\in\mathcal{A}(n_i-1)} \Pr(T\le t|\mathbf{L}_i, A=a, \mathbf{A}_{i,-k}=\mathbf{a}_{i,-k},\hat{\boldsymbol{\omega}}) \\ &\qquad\qquad \times \pi(\mathbf{a}_{i,-k},\alpha) \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_{i,-j}\in\mathcal{A}(n_i-1)} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\hat{\boldsymbol{\omega}}) \\ &= \frac{1}{m}\sum_{i=1}^{m} m_i^{int}(t,a,\alpha,\hat{\boldsymbol{\omega}}) \end{aligned}$$

where $m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\hat{\boldsymbol{\omega}}) = \Pr(T\le t|\mathbf{L}_i, A=a, \mathbf{A}_{i,-k}=\mathbf{a}_{i,-k},\hat{\boldsymbol{\omega}})$. Similarly, the parametric g formula estimator for $\mu(t,\alpha)$ is given as-

$$m^{int}(t,\alpha,\hat{\boldsymbol{\omega}}) = \int \sum_{\mathbf{a}_i\in\mathcal{A}(n_i)} \Pr(T\le t|\mathbf{L}=\mathbf{l}, \mathbf{A}_i=\mathbf{a}_i,\hat{\boldsymbol{\omega}}) dF_{\mathbf{L}}(\mathbf{l})\pi(\mathbf{a}_i,\alpha)$$

and empirically the estimator is calculated as-

$$\begin{aligned} m^{int}(t,\alpha,\hat{\boldsymbol{\omega}}) = &\quad \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_i\in\mathcal{A}(n_i)} \Pr(T\le t|\mathbf{L}_i, \mathbf{A}_i=\mathbf{a}_i,\hat{\boldsymbol{\omega}})\pi(\mathbf{a}_i,\alpha) \\ = &\quad \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_i\in\mathcal{A}(n_i)} m_{ij}(\mathbf{a}_i,t,\mathbf{L}_i,\hat{\boldsymbol{\omega}}) \\ = &\quad \frac{1}{m}\sum_{i=1}^{m} m_i^{int}(t,\alpha,\hat{\boldsymbol{\omega}}) \end{aligned}$$

where $m_{ij}(\mathbf{a}_i,t,\mathbf{L}_i,\hat{\boldsymbol{\omega}}) = \Pr(T\le t|\mathbf{L}_i, \mathbf{A}_i=\mathbf{a}_i,\hat{\boldsymbol{\omega}})$. Based on these parametric g estimators, the causal effect estimates are given by $\widehat{DE}(t,\alpha) = m^{int}(t,0,\alpha,\hat{\boldsymbol{\omega}}) - m^{int}(t,1,\alpha,\hat{\boldsymbol{\omega}})$, $\widehat{IE}(t,\alpha_1,\alpha_2) = m^{int}(t,0,\alpha_1,\hat{\boldsymbol{\omega}}) - m^{int}(t,0,\alpha_2,\hat{\boldsymbol{\omega}})$, $\widehat{TE}(t,\alpha_1,\alpha_2) = m^{int}(t,0,\alpha_1,\hat{\boldsymbol{\omega}}) - m^{int}(t,1,\alpha_2,\hat{\boldsymbol{\omega}})$ and $\widehat{OE}(t,\alpha_1,\alpha_2) = m^{int}(t,\alpha_1,\hat{\boldsymbol{\omega}}) - m^{int}(t,\alpha_2,\hat{\boldsymbol{\omega}})$. Under the assumptions discussed before and when the outcome model is correctly specified, the parametric g formula estimator can be shown to be consistent and asymptotically normal.

4.2.4 Proposed Estimator

An important intrinsic assumption for both the IPCW estimator and the parametric g formula is that the underlying models are specified correctly. If this does not hold true then the consistency and asymptotic normality of the estimators are not valid. The doubly robust estimator provides an alternative to these methods where the estimator is supposedly robust under some model misspecifications. Liu et al. (2018) extended doubly robust estimators to the case of partial interference. The estimators put forth by them has the following form-

$$\hat{F}^{LAN}(a,\alpha) = m^{-1}\sum_{i=1}^{m} n_i^{-1}\sum_{j=1}^{n_i}\Bigg[\frac{I(A_{ij}=a)\{Y_{ij}-m_{ij}(\mathbf{A}_i,\mathbf{L}_i,\hat{\boldsymbol{\omega}})\}\pi(\mathbf{A}_{i,-j};\alpha)}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\hat{\boldsymbol{\beta}})}$$

$$+\sum_{\mathbf{a}_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},\mathbf{L}_i,\hat{\boldsymbol{\omega}})\pi(\mathbf{a}_{i,-j};\alpha)\Bigg]$$

where Y_{ij} is the potential outcome of interest and $m_{ij}(\mathbf{a}_i,\mathbf{L}_i,\boldsymbol{\omega}) = E(Y_{ij}|\mathbf{a}_i,\mathbf{L}_i,\boldsymbol{\omega})$. Extending Liu et al.'s estimate for data with right censoring, considering $Y_{ij} = I(X_{ij}\le t)$ the proposed doubly robust estimate for $\mu(t,a,\alpha)$ is given by

$$\begin{aligned}&\hat{F}^{DR}(t,a,\alpha)\\ &= m^{-1}\sum_{i=1}^{m} n_i^{-1}\sum_{j=1}^{n_i}\Bigg[\frac{I(A_{ij}=a)I(\Delta_{ij}^t=1)\{I(X_{ij}\le t)-m_{ij}(\mathbf{A}_i,t,\mathbf{L}_i,\hat{\boldsymbol{\omega}})\}\pi(\mathbf{A}_{i,-j};\alpha)}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\hat{\boldsymbol{\beta}})\Pr(\Delta_{ij}^t=1|\mathbf{L}_i,X_{ij},\hat{\boldsymbol{\gamma}})}\\ &\qquad\qquad +\sum_{a_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\hat{\boldsymbol{\omega}})\pi(\mathbf{a}_{i,-j};\alpha)\Bigg].\end{aligned}$$

Along the same lines, the doubly robust estimator for $\mu(t,\alpha)$ adjusting for censoring is given by-

$$\hat{F}^{DR}(t,\alpha) = m^{-1}\sum_{i=1}^{m} n_i^{-1}\sum_{j=1}^{n_i}\Bigg[\frac{I(\Delta_{ij}^t=1)\{I(X_{ij}\le t)-m_{ij}(\mathbf{A}_i,t,\mathbf{L}_i,\hat{\boldsymbol{\omega}})\}\pi(\mathbf{A}_i;\alpha)}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\hat{\boldsymbol{\beta}})\Pr(\Delta_{ij}^t=1|\mathbf{L}_i,X_{ij},\hat{\boldsymbol{\gamma}})}$$

$$+\sum_{a_i} m_{ij}(\mathbf{a}_{i,-j}, t, \mathbf{L}_i, \hat{\boldsymbol{\omega}})\pi(\mathbf{a}_i;\alpha)\Big],$$

where $\Delta_{ij}^t = 1$ if $T_{ij}^t(\mathbf{A}_i) \leq C_{ij}$ and $\Delta_{ij}^t = 0$ otherwise and $T_{ij}^t(.) = min(T_{ij}(.), t)$. Using the time dependent censoring indicator Δ_{ij}^t instead of the original censoring indicator Δ_{ij} yields a more efficient estimator in this case because individuals censored after time t will contribute the information in the estimator. And as before, the estimates for the causal effects can be given as follows- $\widehat{DE}(t,\alpha) = \hat{F}^{DR}(t,0,\alpha) - \hat{F}^{DR}(t,1,\alpha)$, $\widehat{IE}(t,\alpha_1,\alpha_2) = \hat{F}^{DR}(t,0,\alpha_1) - \hat{F}^{DR}(t,0,\alpha_2)$, $\widehat{TE}(t,\alpha_1,\alpha_2) = \hat{F}^{DR}(t,0,\alpha_1) - \hat{F}^{DR}(t,1,\alpha_2)$ and $\widehat{OE}(t,\alpha_1,\alpha_2) = \hat{F}^{DR}(t,\alpha_1) - \hat{F}^{DR}(t,\alpha_2)$.

4.2.5 Properties of the Proposed Estimator

The definition of the doubly robust estimator entails the estimation of three sets of nuisance parameter $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\omega}$. For observational data, these parameters are not known beforehand and must be estimated from the observed data to calculate the value of the estimate.

Following the parametric frailty models formulated by Munda et al. (2012), the censoring times c_{ij} were assumed to follow a parametric frailty model. The conditional hazard for the censoring times is given by $g_{ij}(c|\mathbf{L}_{ij}, e_i) = g_0(\boldsymbol{c;\theta_h})e_i \exp(\mathbf{L}_{ij}^T\boldsymbol{\theta_c})$. Here, the baseline hazard function is denoted by g_0, $\boldsymbol{\theta_h}$ is the q'- dimensional vector of parameters for the baseline hazard function, the random effect e_i is assumed to follow density $f_e(e_i;\theta_r)$, and the vector of parameters corresponding to the covariates are denoted by $\boldsymbol{\theta_c}$, which is is q-dimensional. So, the overall vector of parameters for the parametric frailty model $\boldsymbol{\gamma}$ is given as $\boldsymbol{\gamma} = (\boldsymbol{\theta_c}, \boldsymbol{\theta_h}, \theta_r)$. The vector of parameters $\boldsymbol{\gamma}$ is estimated using maximum likelihood estimators. Hence, following assumption III, the contribution of group i to the marginal log-likelihood is (Munda et al. 2012)

$$l(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\gamma}) = \sum_{j=1}^{n_i} \Delta_{ij}^t \left[\log\{g_0(X_{ij}^t)\} + \mathbf{L}_{ij}^T\boldsymbol{\theta_c}\right] + (-1)^{d_i^*}\mathcal{L}^{(d_i^*)} \sum_{j=1}^{n_i} G_0(X_{ij}^t)\exp(\mathbf{L}_{ij}^T\boldsymbol{\theta_c}).$$

Here, $X_{ij}^t = \min(T_{ij}^t(\mathbf{A}_i), C_{ij})$, $\mathbf{X}_i^t = (X_{i1}^t, X_{i2}^t, \cdots, X_{in_i}^t)$, $\mathbf{\Delta}_i^t = (\Delta_{i1}^t, \Delta_{i2}^t, \cdots, \Delta_{in_i}^t)$, $d_i^t = \sum_{j=1}^{n_i}(1 - \Delta_{ij}^t)$ is the number of censored observations in group i at time t, $G_0(\omega) = \int_0^\omega g_0(\kappa)d\kappa$, and $\mathcal{L}^{(s)}$ is the s-th derivative of the Laplace transform of the frailty distribution, i.e., $\mathcal{L}^{(s)} = \int_0^\infty \exp(-e_i s) f_e(e_i; \theta_r) de_i$. In terms of estimating equation theory, the maximum likelihood estimator of $\boldsymbol{\gamma}$ can be formulated as solution of the following estimating equations

$$\sum_i \psi_{ck}(\mathbf{X}_i, \mathbf{\Delta}_i, \mathbf{L}_i, \boldsymbol{\gamma}) = 0 \text{ for } k = 1, ..., q + q' + 1,$$

where $\psi_{ck} = \psi_{ck}(\mathbf{X}_i, \mathbf{\Delta}_i, \mathbf{L}_i, \boldsymbol{\gamma}) = \partial l(\mathbf{X}_i^t, \mathbf{\Delta}_i^t, \mathbf{L}_i, \boldsymbol{\gamma})/\partial\gamma_k$ and γ_k is the k-th element of $\boldsymbol{\gamma}$.

The calculation of censoring weight can be simplified by the choice of baseline hazard and frailty distribution. Specifically, in this section, a constant value equal to θ_h is used as the baseline hazard for the censoring model.Then the conditional hazard is given as $g_{ij}(c|\mathbf{L}_{ij}, e_i) = \theta_h \exp(\mathbf{L}_{ij}\boldsymbol{\theta_c})e_i$. Also, it is assumed that the frailty term e_i follows a gamma distribution with mean 1 and variance θ_r. Under these specifications, the censoring weights can be obtained exactly using the closed form obtained below-

$$\begin{aligned}
\Pr(\Delta_{ij}^t = 1|\mathbf{L}_i, X_{ij}, \boldsymbol{\gamma}) &= \int \Pr(C_{ij} > X_{ij}^t|\mathbf{L}_i, \boldsymbol{\gamma}, e_i) f_e(e_i; \theta_r) de_i \\
&= \int \exp\{-\theta_h X_{ij}^t \exp(\mathbf{L}_{ij}\boldsymbol{\theta_c})e_i\} \frac{e_i^{1/\theta_r - 1} e^{-e_i/\theta_r}}{\theta_r^{1/\theta_r}\Gamma 1/\theta_r} de_i \\
&= \left\{\frac{1}{\theta_r\theta_h X_{ij}^t \exp(\mathbf{L}_{ij}\boldsymbol{\theta_c}) + 1}\right\}^{1/\theta_r}
\end{aligned}$$

Next, the group propensity weights must be estimated. TV (2012), used a mixed effects model for the treatment indicator, i.e., $\Pr(A_{ij} = 1|\mathbf{L}_{ij}, b_i) = \text{logit}^{-1}(\mathbf{L}_{ij}\boldsymbol{\theta_x} + b_i)$ to estimate the group propensity scores. Here, $\boldsymbol{\theta_x}$ corresponds to the covariate parameters and b_i is a random effect following density $f_b(b_i; \theta_s)$. Then, the vector of parameters for the mixed effects model is give by $\boldsymbol{\beta} = (\boldsymbol{\theta_x}, \theta_s)$. Maximum likelihood theory is again employed for the estimation of the parameter vector $\boldsymbol{\beta}$. The contribution of group i to

the log-likelihood for the mixed effects model is given by

$$l(\mathbf{A}_i, \mathbf{L}_i, \boldsymbol{\beta}) = \log \left[\int \prod_{j=1}^{n_i} h_{ij}(b_i, \mathbf{L}_i, \boldsymbol{\theta_x})^{A_{ij}} \{1 - h_{ij}(b_i, \mathbf{L}_i, \boldsymbol{\theta_x})\}^{(1-A_{ij})} f_b(b_i; \theta_s) \right],$$

where $h_{ij}(b_i, \mathbf{L}_i, \boldsymbol{\beta}) = \Pr(A_{ij} = 1 | \mathbf{L}_{ij}, b_i)$. The likelihood for group i is obtained by integrating over the random effect for group i. Similarly as before, the maximum likelihood estimator of $\boldsymbol{\beta}$ can be formulated as the solution to the vector of estimating equations

$$\sum_i \psi_{xk}(\mathbf{A}_i, \mathbf{L}_i, \boldsymbol{\beta}) = 0 \text{ for } k = 1, ..., p+1,$$

where $\psi_{xk} = \psi_{xk}(\mathbf{A}_i, \mathbf{L}_i, \boldsymbol{\beta}) = \partial l(\mathbf{A}_i, \mathbf{L}_i, \boldsymbol{\beta}) / \partial \beta_k$, β_k is the k-th element of $\boldsymbol{\beta}$.

Finally, the time to events must be modeled. According to Munda et al. (2012), again a parametric frailty model can be employed for this purpose. For the outcome model, the conditional hazard $g_{ij}^{OR}(t | \mathbf{L}_{ij}, A_{ij}, \mathbf{A}_{i,-j}, e_i^{OR})$ for $T_{ij}(\mathbf{a}_i)$ is given by

$$g_{ij}^{OR}(t | \mathbf{L}_{ij}, A_{ij}, \mathbf{A}_{i,-j}, \boldsymbol{\omega}, e_i^{OR}) = g_0^{OR}(\boldsymbol{\omega}_h, t) e_i^{OR} \times \exp\left(\mathbf{L}_{ij}^T \boldsymbol{\omega}_{c(1:p_1')} + \phi(\mathbf{A}_{i,-j}) \boldsymbol{\omega}_{c(p_1'+1:p')} + A_{ij} \omega_{c(p'+1)}\right).$$

Again, as before, g_0^{OR} denotes the baseline hazard function, the p'' dimensional parameter vector corresponding to the baseline hazard function is given by $\boldsymbol{\omega}_h$, e_i^{OR} is the frailty term following density $f_e^{OR}(e_i^{OR}; \omega_r)$, $\boldsymbol{\omega}_c = (\boldsymbol{\omega}_{c(1:p_1')}, \boldsymbol{\omega}_{c(p_1'+1:p')}, \omega_{c(p'+1)})$ is a $p'+1$ dimensional vector of coefficients corresponding to the covariates and ω_r is a parameter of the random effect model. The overall vector of parameters corresponding to the outcome model is then $\boldsymbol{\omega} = (\boldsymbol{\omega}_c, \boldsymbol{\omega}_h, \omega_r)$. This vector of nuisance parameter is not known beforehand and must be estimated from the data. The treatment assignment vector is incorporated in the model through the function ϕ. For example, in this section, ϕ is assumed to be the proportion of treated individuals in each group. As before, the contribution of group i, due to assumption III, towards the marginal log-likelihood equals (Berg and Drepper

2012)

$$
\begin{aligned}
& l(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\omega}) \\
& \quad = \left(\sum_{j=1}^{n_i} \Delta_{ij} \left[\log\{g_0^{OR}(X_{ij})\} + \mathbf{L}_{ij}^T \boldsymbol{\omega}_{c(1:p_1')} + \phi(\mathbf{A}_{i,-j}) \boldsymbol{\omega}_{c(p_1'+1:p')} + A_{ij} \omega_{c(p'+1)} \right] \right) + \\
& \quad \left[(-1)^{d_i} \mathcal{L}^{(d_i)} \left\{ \sum_{j=1}^{n_i} g_0^{OR}(X_{ij}) \exp\left(\mathbf{L}_{ij}^T \boldsymbol{\omega}_{c(1:p_1')} + \phi(\mathbf{A}_{i,-j}) \boldsymbol{\omega}_{c(p_1'+1:p')} + A_{ij} \omega_{c(p'+1)} \right) \right\} \right].
\end{aligned}
$$

Here $d_i = \sum_{j=1}^{n_i} \Delta_{ij}$ denotes the number of uncensored cases, $\mathcal{L}^{(q)}$ denotes the q-th derivative of the Laplace transform of the frailty distribution, i.e.,

$$
\mathcal{L}^{(s)} = \int_0^{\infty} \exp\left(-e_i^{OR} s\right) f(e_i^{OR}, \omega_r) de_i^{OR}, \qquad s \geq 0.
$$

The parameters are estimated using maximum likelihood. For obtaining these, the corresponding score equations must be maximized. Hence, these score equations can be represented as a set of estimating equations and the maximum likelihood estimate for $\boldsymbol{\omega}$ can be interpreted as the solution of the following estimating equations

$$
\sum_i \psi_{ck}^{OR}(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\omega}) = 0 \text{ for } k = 1, ..., p' + p'' + 1,
$$

where $\psi_{ck}^{OR} = \psi_{ck}^{OR}(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\omega}) = \partial l(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\omega}) / \partial \omega_k$, ω_k being the k-th member of $\boldsymbol{\omega}$. For calculating the estimators proposed in Section 4.2.4, quantity $\Pr(T \leq t | \mathbf{L}_i, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}})$ must be calculated. Using the estimated values of $\hat{\boldsymbol{\omega}}$, the term can be calculated easily and the proposed estimator can be obtained by integrating out the effect of the random component. If the estimated survival function corresponding to the

model for group i is given by $S_g(g_{ij}^{OR}(t|\mathbf{L}_i, A_{ij}, \mathbf{A}_{i,-j}, e_i^{OR}, \hat{\boldsymbol{\omega}}))$, then,

$$\Pr(T \leq t|\mathbf{L}_i, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}}) = 1 - \int S_g(g_{ij}^{OR}(t|\mathbf{L}_i, A_{ij} = a, \mathbf{A}_{i,-j}, e_i^{OR}, \hat{\boldsymbol{\omega}}))f_e^{OR}(e_i^{OR}; \omega_r)de_i^{OR}$$

So, to estimate the probability $\Pr(T \leq t|\mathbf{L}_i, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}})$, a parametric frailty model is assumed for the time to events and the conditional survival probabilities must be integrated with respect to the random component. In all the analysis performed in this section, the parametric frailty model employed is assumed to have a frailty distribution. of gamma. Specifically, e_i^{OR} follows gamma distribution with variance ω_r. For a gamma frailty, the quantity $\Pr(T \leq t|\mathbf{L}_i, A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}})$ can be obtained as a closed form by computing the exact integral mathematically. Using integral calculus we can show the following-

$$\begin{aligned}\Pr(T \leq t|\mathbf{L}_i, &A = a, \mathbf{A}_{i,-k} = \mathbf{a}_{i,-k}, \hat{\boldsymbol{\omega}}) \\ &= 1 - \int \exp\{-G_0^{OR}(\mathbf{L}_i, \mathbf{A}_{i,-j}, a, \hat{\boldsymbol{\omega}}, t)e_i^{OR}\} \frac{(e_i^{OR})^{1/\hat{\omega}_r - 1} e^{-e_i^{OR}/\hat{\omega}_r}}{\hat{\omega}_r{}^{1/\hat{\omega}_r}\Gamma 1/\hat{\omega}_r} de_i^{OR} \\ &= 1 - \left\{\frac{1}{\hat{\omega}_r G_0^{OR}(\mathbf{L}_i, \mathbf{A}_{i,-j}, a, \hat{\boldsymbol{\omega}}, t) + 1}\right\}^{1/\hat{\omega}_r}\end{aligned}$$

Here,

$$G_0^{OR}(\mathbf{L}_i, \mathbf{A}_{i,-j}, a, \hat{\boldsymbol{\omega}}, t) = \exp\left(\mathbf{L}_{ij}^T \hat{\boldsymbol{\omega}}_{c(1:p_1')} + \phi(\mathbf{A}_{i,-j})\hat{\boldsymbol{\omega}}_{c(p_1'+1:p')} + a\hat{\omega}_{c(p'+1)}\right) \times \int_0^t g_0^{OR}(\hat{\boldsymbol{\omega}}_h, s)ds$$

So, it follows that

$$m_i^{int}(t,a,\alpha,\hat{\boldsymbol{\omega}}) = 1 - \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_{i,-j}\in\mathcal{A}(n_i-1)}\left\{\frac{1}{\hat{\omega_r}G_0^{OR}(\mathbf{L}_i,\mathbf{A}_{i,-j},a,\hat{\boldsymbol{\omega}},t)+1}\right\}^{1/\hat{\omega_r}}\pi(\mathbf{a}_{i,-k},\alpha)$$

and

$$m^{int}(t,a,\alpha,\hat{\boldsymbol{\omega}}) = \frac{1}{m}\sum_{i=1}^{m} m_i^{int}(t,a,\alpha,\hat{\boldsymbol{\omega}}).$$

Similarly,

$$m_i^{int}(t,\alpha,\hat{\boldsymbol{\omega}}) = 1 - \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_i\in\mathcal{A}(n_i)}\sum_{a'=0}^{1}\left\{\frac{1}{\hat{\omega_r}G_0^{OR}(\mathbf{L}_i,\mathbf{A}_{i,-j},a',\hat{\boldsymbol{\omega}},t)+1}\right\}^{1/\hat{\omega_r}}\pi(\mathbf{a}_i,\alpha)$$

and

$$m^{int}(t,\alpha,,\hat{\boldsymbol{\omega}}) = \frac{1}{m}\sum_{i=1}^{m} m_i^{int}(t,\alpha,,\hat{\boldsymbol{\omega}})$$

Therefore, considering the estimation of all the aforementioned parameters as well as the parameters of interest for obtaining the causal effects, the overall vector of estimating equations are obtained by stacking all of the estimating equations and can be denoted by

$$\sum_i \psi^{DR}(\mathbf{O}_i,\boldsymbol{\theta}^{DR}) = 0,$$

where, $\mathbf{O}_i = (\mathbf{X}_i,\boldsymbol{\Delta}_i,\mathbf{A}_i,\mathbf{L}_i)$, $\psi(\mathbf{O}_i,\boldsymbol{\theta}^{DR}) = \left(\boldsymbol{\psi_c},\boldsymbol{\psi_x},\boldsymbol{\psi}_c^{OR},\psi_{a\alpha}^{DR}\right)^T$, $\boldsymbol{\psi_c} = (\psi_{c1},\psi_{c2},...,\psi_{cq+q'+1})^T$, $\boldsymbol{\psi_x} = (\psi_{x1},\psi_{x2},...,\psi_{xp})^T$, $\boldsymbol{\psi}_c^{OR} = \left(\psi_{c1}^{OR},\psi_{c2}^{OR},...,\psi_{c(p'+p''+1)}^{OR}\right)^T$, and $\psi_{a\alpha}^{DR} = \psi_{a\alpha}^{DR}(\mathbf{O}_i,\boldsymbol{\theta}^{DR}) = \hat{F}^{DR}(t,a,\alpha) - \theta_{a\alpha}^{DR}$. The vector of parameters to be estimated from the estimating equations is $\boldsymbol{\theta}^{DR} = \left(\boldsymbol{\theta}_c,\boldsymbol{\theta_h},\theta_r,\boldsymbol{\theta_x},\theta_s,\boldsymbol{\omega},\theta_{a\alpha}^{DR}\right)$. The overall vector of solutions are obtained using maximum likelihood theory and using the proposed estimators and is given by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_c,\hat{\boldsymbol{\theta}}_{\boldsymbol{h}},\hat{\theta}_r,\hat{\boldsymbol{\theta}}_{\boldsymbol{x}},\hat{\theta}_s,\hat{\boldsymbol{\omega}},\hat{\theta}_{a\alpha}^{DR})$. Finally, the true value of the parameters are denoted by of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\omega}$ by $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_0$, and $\boldsymbol{\omega}_0$.

Proposition 3. *If either (i) the propensity model and the censoring model are correctly specified or (ii) the outcome model is correctly specified then* $\sqrt{m}\{\hat{F}^{DR}(t,a,\alpha) - \mu(t,a,\alpha)\}$ *converges to a normal distribution with mean* 0 *and variance* $\boldsymbol{\Sigma}^{DR}$ *as* $m \to \infty$ *where* $\boldsymbol{\Sigma}^{DR} = \tau U(\boldsymbol{\theta}^{DR})^{-1} V(\boldsymbol{\theta}^{DR})\{U(\boldsymbol{\theta}^{DR})^{-1}\}^T \tau^T$ *where* $U(\boldsymbol{\theta}^{DR}) = E\{-\dot{\psi}^{DR}(\mathbf{O}_i, \boldsymbol{\theta}^{DR})\}$, $V(\boldsymbol{\theta}^{DR}) = E\{\psi^{DR}(\mathbf{O}_i, \boldsymbol{\theta}^{DR})\psi^{DR}(\mathbf{O}_i, \boldsymbol{\theta}^{DR})^T\}$, $\dot{\psi}^{DR}(\mathbf{O}_i, \boldsymbol{\theta}^{DR}) = \partial\psi^{DR}(\mathbf{O}_i, \boldsymbol{\theta}^{DR})/\partial(\boldsymbol{\theta}^{DR})^T$, *and* $\tau = (0, 0, \ldots, 0, 1)$.

The proof of Proposition 3 is given in Section 4.6. The asymptotic variance $\boldsymbol{\Sigma}^{DR}$ can be consistently estimated by $\hat{\boldsymbol{\Sigma}}^{DR} = \hat{U}(\hat{\boldsymbol{\theta}}^{DR})^{-1}\hat{V}(\hat{\boldsymbol{\theta}}^{DR})\{\hat{U}(\hat{\boldsymbol{\theta}}^{DR})^{-1}\}^T$ where $\hat{U}(\hat{\boldsymbol{\theta}}^{DR}) = m^{-1}\sum_{i=1}^{m}\{-\dot{\psi}^{DR}(\mathbf{O}_i, \hat{\boldsymbol{\theta}}^{DR})\}$ and $\hat{V}(\hat{\boldsymbol{\theta}}^{DR}) = m^{-1}\sum_{i=1}^{m}\{\psi^{DR}(\mathbf{O}_i, \hat{\boldsymbol{\theta}}^{DR})\psi^{DR}(\mathbf{O}_i, \hat{\boldsymbol{\theta}}^{DR})^T\}$. The sandwich variance estimator $\hat{\boldsymbol{\Sigma}}^{DR}$ can be computed using the R package `geex` (Saul and Hudgens 2017) and can be used to construct Wald type confidence intervals.

4.3 Simulation

To demonstrate the efficacy of the methods discussed, a simulation study was performed. The data was simulated using the following steps

1. First, baseline covariates L_{1ij}, L_{2ij}, L_{3ij}, and L_{4ij} was generated randomly as follows. L_{1ij} was generated following a standard normal distribution, L_{2ij} was generated following a Bernoulli distribution with probability of success 0.5, and L_{3ij} was generated following a chi-square distribution with 1 degree of freedom. To generate L_{4ij}, first, random variables V_{ij} were generated following an inverse Gaussian distribution with mean 1. The baseline covariates L_{4ij} were then assigned to be $V_{ij}^2 + L_{3ij}$.

2. Random effects b_i for generating the treatment probabilities were randomly generated following a normal distribution with mean 0 and variance 1.

3. Treatment indicators were simulated following a Bernoulli distribution with probability p_{ij} where $p_{ij} = \text{expit}(1 - \beta|L_{1ij}| + b_i)$. The value of β is varied from 0.1 to 1.0

with an increment of 0.1 to get 10 sets of simulated data. The different simulation settings are denoted by alphabets a through j respectively.

4. Random components e_i^{OR} for generating the potential times to event were generated randomly from a gamma distribution with mean 1 and variance 0.1.

5. Then the time to events $T_{ij}(\mathbf{a}_i)$ were randomly sampled from an exponential distribution with mean μ_{ij} where $1/\mu_{ij} = 0.6e_i^{OR}\exp\left(-7.4a_{ij} - 12.7\sum_{k\neq j} a_{ik}/n_i + 10|L_{1ij}| - 7.4L_{2ij} + 20L_{1ij}L_{2ij}\right)$.

6. Next, another set of random effects e_i were generated randomly following a gamma distribution with mean 1 and variance 10

7. Finally, the censoring times C_{ij} were sampled randomly from an exponential distribution with mean λ_0 where where $1/\lambda_0 = 0.01e_i\exp\left(2|L_{1ij}| - 10L_{2ij} + 5L_{1ij}L_{2ij}\right)$.

8. An individual was censored, i.e. $\Delta_{ij} = 1$ if $C_{ij} < T_{ij}(\mathbf{A}_i)$.

9. The misspecified outcome model was obtained by fitting a parametric frailty model to the time to events with the covariates treatment, $\sum_{k\neq j} a_{ik}/n_i$, $|L_{1ij}|$, and L_{2ij}.

10. The misspecified treatment model was obtained by fitting a logistic mixed effects model to the treatment indicators with covariates L_{4ij}.

11. Finally, the misspecified censoring model was obtained by fitting a parametric frailty model to the censoring times with covariate L_{1ij}.

The steps from 1 to 8 were performed iteratively a large number of times. For each of the simulated data, the number of groups was fixed to be 200 and the number of individuals in each group was fixed to be 30. So, for every set of simulation, the total number of individuals was 6000. The true values of the parameters of interest were obtained from the simulated data by using the complete simulated data on the counterfactuals for a large data set. The allocation probabilities used ranged from 0.1 to 0.9 with an increment

of 0.1. So, the total number of allocation probabilities explored was 9. The estimates of the parameters of interest were also obtained for each of the simulated data set. These estimates were compared with the true value of the estimates. Note that the Bernoulli allocation probabilities, in this case, have the following property: if $\sum_{k\neq j} a_{ik}$ is the same for two different values of the vector $\mathbf{a}_{i,-j}$, then $\pi(\mathbf{a}_{i,-j}, \alpha)$ will also be the same for those two treatment vectors. Hence, in this case, the population average potential cumulative survival distribution at the time point t for treatment a under allocation strategy α can be simplified as (Perez-Heydrich et al. 2014)

$$m^{-1}\sum_i n_i^{-1}\sum_j\sum_{k=0}^{n_i-1}\binom{n_i-1}{k}I\{T_{ij}(a,\mathbf{a}_{i,-j})\leq t\}\alpha^k(1-\alpha)^{n_i-k-1}$$

An equivalent explanation yields that the population average marginal potential cumulative survival distribution at the time point t under allocation strategy α can be represented as

$$m^{-1}\sum_i n_i^{-1}\sum_j\sum_{k=0}^{n_i}\binom{n_i}{k}I\{T_{ij}(a,\mathbf{a}_{i,-j})\leq t\}\alpha^k(1-\alpha)^{n_i-k-1}$$

This is a special case of partial interference and this representation is possible because of the fact that the contribution of the vector $\mathbf{a}_{i,-j}$ towards the time to events is only through the sum of its elements $\sum_{k\neq j} a_{ik}$. The estimated asymptotic standard errors were also obtained using the sandwich variance estimator discussed in the previous section for each of the simulated data for each α. The representative value of the estimate and the asymptotic standard error were calculated as the average over all the simulated data set. Standard errors were also calculated empirically from the estimates obtained from each of the simulated data set. The asymptotic standard errors were used to construct 95% Wald confidence intervals. Finally, for each value of α, coverage percentages were calculated by observing the percentage of estimated values lying within the confidence interval over all the simulated data set.

Fitting all the correctly specified models, 1000 iterations of simulation setting e were performed. The results are summarized in Table 4.1, which describes the results obtained in detail for nine different α values at time point 100 days. From the table, it can be observed that the bias is close to 0 in general. Also, the Wald type 95% confidence intervals seem to attain the nominal level.

Figure 4.1 and Figure 4.2 show the absolute biases of the estimates $\hat{F}^{DR}(t, a, \alpha)$ for the different simulation settings a through j for untreated and treated individuals respectively for $\alpha = 0.5$. Similarly, Figure 4.3 and Figure 4.4 show the coverages of the 95% Wald type confidence intervals for the different simulation settings a through j for untreated and treated individuals respectively for $\alpha = 0.5$. These plots contain results for IPCW, parametric g as well as doubly robust estimators under four different scenarios. These are, i) when the outcome model, treatment model, and censoring model are all specified correctly, ii) when the outcome model is specified incorrectly but the treatment and censoring models were specified correctly, iii) when only the outcome model was specified correctly but both the treatment and censoring model were specified correctly, and iv) when all the three models are specified incorrectly. The IPCW estimators were particularly sensitive towards the actual distribution of the propensity scores. Depending on how the propensity scores were distributed, the IPCW estimators were either always unbiased or always biased irrespective of the models being correctly or incorrectly specified. Simulation scenarios a through j were used to compare the estimators for different propensity distributions. From the figures, note that the bias of the IPCW method for incorrectly specified treatment and censoring models is quite close to 0 for simulation setting a and increases from simulation setting b through j. However, if only one set of models are incorrectly specified, then, across all the simulations, the doubly robust estimator performs uniformly better in terms of bias as well as coverage than the corresponding estimator for which the model(s) has(have) been misspecified thus demonstrating the doubly robust property of the estimator.

4.4 Data Analysis

4.4.1 Cholera Vaccine Study and Analysis

In order to show the performance of the methods described in Section 4.2, data from a cholera vaccine study in Matlab, Bangladesh (Ali et al. 2005) are analyzed. The study participants consisted of children of ages 2-15 and women. There were two versions of vaccine administered. They were B subunit-killed whole-cell oral cholera vaccine or killed whole-cell-only cholera vaccine. The placebo used on the participants for this study was E. coli K12 placebo. However, according to SUTVA, it is assumed that there is only one version of treatment and control. Hence, the two treatments are considered to be the same treatment. Individuals in the study randomly received one of the three treatments. Participation in the study was based on whether an individual received two or more doses of the treatment or vaccine assigned to him/her. Since non participants were also included in the analysis of the study, a matrix with information on whether an individual participated in the study or not was maintained. The assignment of treatment or control to an individual in the Matlab study was irrespective of their participation status. The dates of vaccination ranged from January to May 1985. Three centers for vaccination were established in the Matlab area and later, these were used as surveillance centers for end point data collection. The total number of individuals in the data consisted was 121,982. The presence of interference has been established by a number of studies of the data in the past (Perez-Heydrich et al. 2014). However, in all of those studies, the issue of right censoring was ignored.

Not all the individuals in the study observe an event of cholera. For those who do not, the time to incidence of cholera is censored. Hence, the presence of right censoring is evident in the data. There are a number of causes of censoring like migration from study location or death during the follow up period.

Before proceeding with the analysis, the data was prepared accordingly as follows.

7 duplicate values were observed in the 121,982 observed individuals. These observations were removed from the data and there remained a total of 121,975 individuals. Note that, the date of vaccination was not the same for all individuals as previously mentioned. However, since the range of vaccination was only five months, the date of vaccination of individuals were quite close to each other. Therefore, for convenience of analysis, individuals within the same group were assigned the same date of entry into the study. The entry date for an individual was defined to be the most recent date of second vaccination among all the individuals within the group to which that particular individual belonged. There were some groups in which none of the individuals received a second vaccination. For those groups, the start date was defined as the most recent date of first vaccination among all the individuals within that group. During the process of assigning start dates to individuals, it was observed that 34 individuals could not be assigned any start date as nobody in their group were given a single dose of either vaccine or placebo. These observations were also removed giving rise to 121,941 individuals. After that, it was again observed that 60 of the observations had start dates assigned to them which were before the date of them contracting cholera. Since time to contract cholera cannot be negative, those observations were deleted. Along similar lines, the dates of migration of 3617 individuals and the dates of death for 346 individuals were were observed to be before their group's start date and hence they had to be removed as well and 117,918 individuals remained. This is the dimension of the final data that was ultimately analyzed. Within this, there were 69,219 individuals who received placebo and there were 48,699 individuals who received treatment. The number of cases of cholera was very small, i.e., in the control and treatment group, there were 375 and 103 cases of cholera reported respectively.

While calculating the propensity score for the estimator, the largest 15 groups posed a problem as their group propensity scores were very close to 0. Because of this, the weights were abnormally large and numerical calculations could not be performed using those

weights. So, these groups had to be be omitted as well. The probability of vaccination was 2/3 in the data if the two groups of vaccines were merged together. The probability of participation was modeled using a mixed effects model following Perez-Heydrich et. al.. Age, squared age, distance to nearest river, and squared distance to nearest river were all included as covariates in the model. The propensity score for group i was estimated by

$$\Pr(\mathbf{A}_i|\mathbf{L}_i,\hat{\boldsymbol{\beta}}) = \int \prod_{j=1}^{n_i}\{(2/3)h_{ij}(b_i,\mathbf{L}_{ij},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\}^{A_{ij}}\{1-(2/3)h_{ij}(b_i,\mathbf{L}_{ij},\hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\}^{(1-A_{ij})} \times f_b(b_i;\hat{\theta}_s),$$

where $h_{ij}(b_i,\mathbf{L}_i,\boldsymbol{\theta}_x)\} = \Pr(B_{ij}=1|b_i,\mathbf{L}_{ij},\boldsymbol{\theta}_x)) = \text{expit}(\mathbf{L}_{ij}\boldsymbol{\theta}_{\boldsymbol{x}} + b_i)$, and $(\hat{\boldsymbol{\theta}}_{\boldsymbol{x}},\hat{\theta}_s)$ is the maximum likelihood estimate of $(\boldsymbol{\theta}_{\boldsymbol{x}},\theta_s)$. For the censoring model, checking individually for statistically significant predictors of censoring, age was chosen as the most significant predictor. Specifically, the censoring times were modeled using a parametric frailty model as exponential random variables with mean $\theta_h c_i \exp(\text{age}_{ij}\theta_c)$, where c_i was assumed to follow a gamma distribution with mean 1 and variance θ_r.

Following Section 4.2, a parametric frailty model is fitted to the outcome of interest, i.e., time to Cholera. For completely specifying the outcome model, there must be a baseline hazard function and a frailty distribution given. Many different combinations of choices were considered for this model.

The AIC and BIC values for several baseline hazard function and frailty distribution combinations are provided in Table 4.2. Both the AIC and BIC values seem to reach their minimum value corresponding to baseline hazard distribution gompertz and frailty distribution gamma. So, the times to incidence of cholera were assumed to follow a parametric frailty model with baseline hazard distribution gompertz and frailty distribution gamma. For gompertz baseline hazard, the function $g_0^{OR}(\boldsymbol{\omega}_h,t)$ equals $\omega_{h2}exp(\omega_{h1}t)$. Treatment status, proportion of individuals treated in the group, interaction of treatment status

and proportion of individuals treated within the group, age, distance to river, squared age and squared distance to river were all used as explanatory variables in the outcome model. Note that, from Section 4.2.5, we can obtain a closed form expression for both the outcome probability as well as the likelihood function since the frailty is gamma. The allocation probabilities that were used for analysis ranged from 0.3 and 0.6 because over 70% of individuals had values of α within that range.

4.4.2 Results

The direct, indirect, total and overall effects at one year per thousand persons obtained from the cholera vaccine study are summarized in Figure 4.5. From the figure, it can be observed that the direct effect is a function of the policy α. The direct effect seems to be inversely related to the allocation probability. Also, the 95% confidence interval for the direct effect includes 0 only for α greater than 0.55. So, for α less than or equal to 0.55, there exists a statistically significant direct effect of vaccine on cholera. To illustrate this, it can be seen that the direct effects and 95% confidence intervals corresponding to $\alpha = 0.45$ and $\alpha = 0.6$ are 4.3 $(1.8, 6.9)$ and 1.6 $(-0.4, 3.7)$ respectively. So, it can be said that in a population of 1000 individuals, 7.8 more individuals are expected to be infected with cholera if they are not vaccinated compared to if they are vaccinated. Unlike the direct effect, all three of the indirect, total and overall effects seem to increase with policy α. All of these three effects provide different measures of the spillover effect and the trend suggests a positive spillover effect of the vaccination on the incidence of cholera. for example, the estimate of the overall effect corresponding to $\alpha_1 = 0.55$ and $\alpha_2 = 0.4$ turns out to be 2.7 with a 95% confidence interval of $(1.3, 4.1)$. This means that almost 3 more individuals are expected to be infected with cholera in a region with allocation probability 0.4 compared to a region with allocation probability 0.55 irrespective of their individual treatment status.

Perez-Heydrich et al. (2014) provided IPW estimates for the direct, indirect, total and overall effects using group propensity scores instead of individual propensity scores. The estimates from the doubly robust estimate compare well with the previous results. The plots of the effects look to be quite similar and the trend is similar as well. For example, the direct effect (and 95% CI) provided by Perez-Heydrich et. al. is given by 4.03 $(2.48, 8.12)$. Whereas the doubly robust method suggested in this section produces a point estimate (and 95% CI) of 7.13 $(4.06, 10.20)$. However, the estimators provided by Perez-Heydrich et. al. suffered from two serious drawbacks that have been addressed in the methods of this section. The first drawback is that for the IPW estimator, the propensity model must be correctly specified. But for the doubly robust method, even if the propensity model and the censoring models are both specified incorrectly, the estimator might be consistent if the outcome model can be specified correctly. Also, the IPW estimators do not adjust for censoring.

4.5 Discussion

In this chapter, the doubly robust estimator proposed by Liu et al. (2018) was extended for data with right censoring. A censoring weight was incorporated in the estimator proposed by them to calculate the new estimator. Using M-estimation theory, the estimator was shown to be consistent and asymptotically normal when either the treatment model and censoring model are correctly specified or only the outcome model is correctly specified, validating its doubly robust property. A sandwich variance estimator was used to construct 95% confidence interval for the estimates. Results from simulation studies showed that the estimator performed well for finite samples, i.e. the estimator was robust under model misspecification, it had a low bias, and the coverage was close to the nominal level of 95%.

The methods were implemented on the cholera vaccine trial and various causal effects

were estimated. Results were compared with previous studies (Ali et al. 2005, Perez-Heydrich et al. 2014). In agreement with these studies, it was observed that there was an inverse relationship of incidence of cholera with the allocation probability. The effects were obtained as a function of both time and policy. However, none of the previous studies adjusted for censoring hence being susceptible to selection bias. (Perez-Heydrich et al. 2014) used an IPW estimator which is sensitive to model misspecification. Results corresponding to time point one year were discussed in detail in this chapter.

The doubly robust estimator suffers from some of the drawbacks of both the IPCW and the parametric g formula. For example, large groups might yield a very small value of group propensity score resulting in an unstable estimator. One way around this issue can be to use standardized estimator instead of the original one proposed in this chapter. The parametric g formula estimate might prove to be mathematically cumbersome to compute as it involves summing over all possible $\mathbf{a}_i$. Liu et al. (2018) used a Monte-Carlo approach to carry out calculations which might be employed n this case as well. A future direction of work might be to adopt a semi parametric Cox proportional hazard model instead of the parametric frailty model for the censoring as well as the outcome model. Naimi and Kennedy (2017) showed that doubly robust methods perform well when used in conjunction with non-parametric models. This is another avenue worth exploring in the future. Finally, the methods discussed here might be extended from partial interference to general interference.

Figure 4.1: Absolute biases of the doubly robust, parametric g and the IPCW estimators under different model misspecifications in control group.

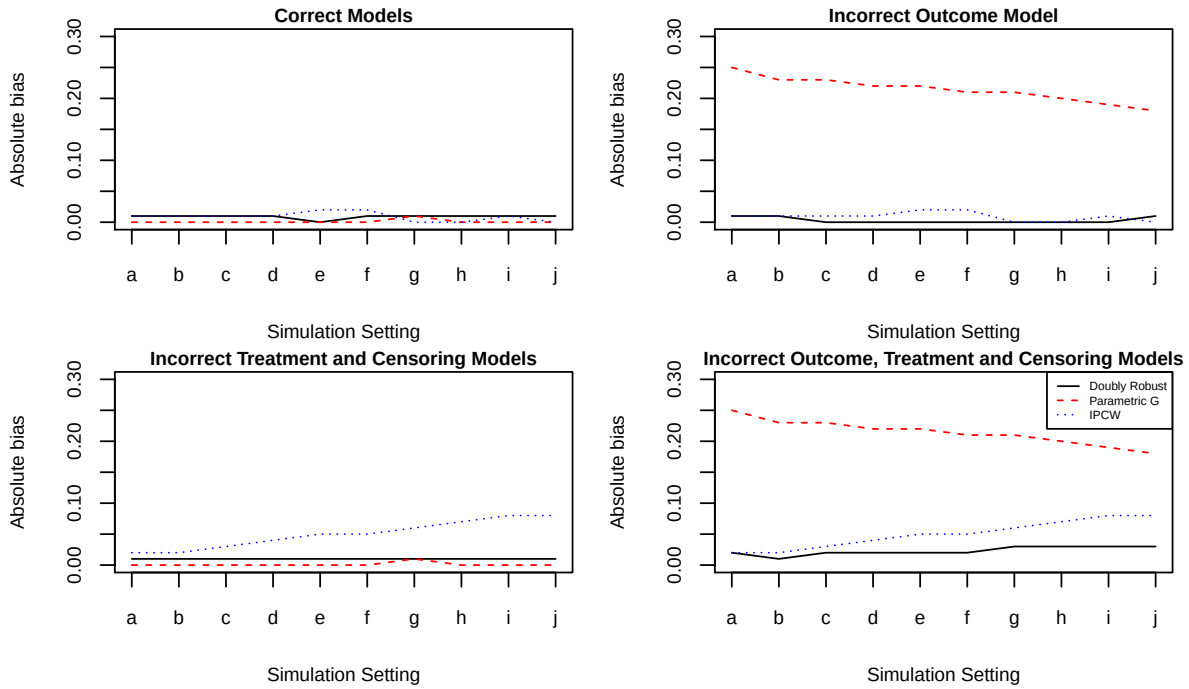


Figure 4.2: Absolute biases of the doubly robust, parametric g and the IPCW estimators under different model misspecifications in treatment group.

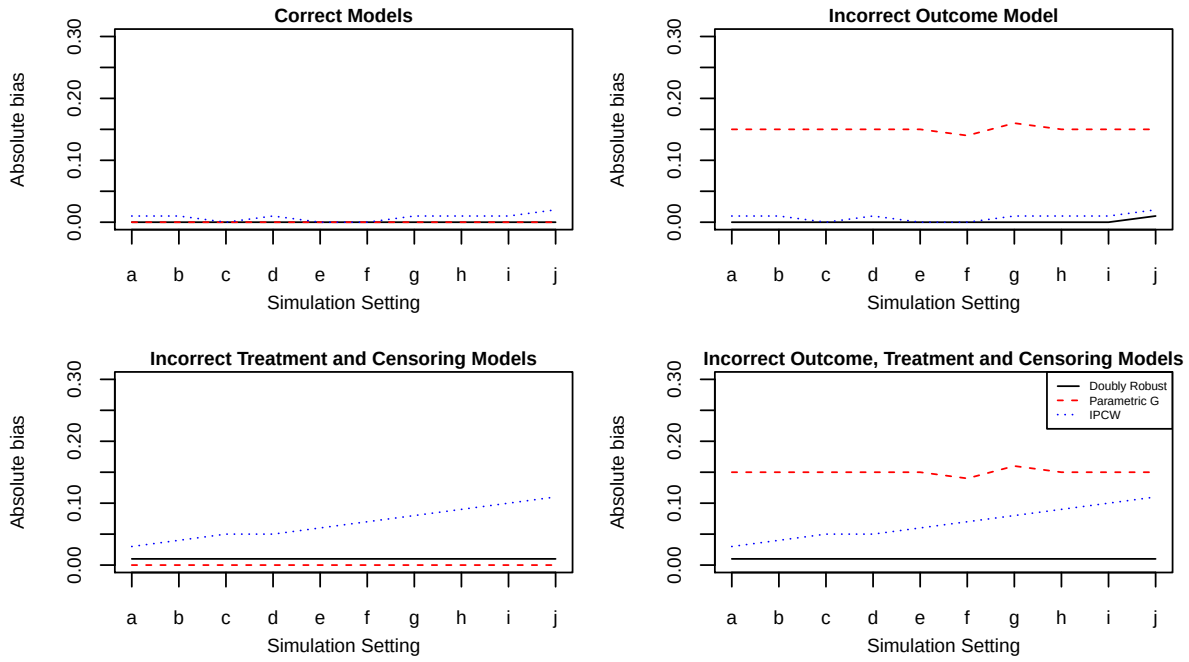


Figure 4.3: Coverages of the doubly robust, parametric g and the IPCW estimators under different model misspecifications in control group.

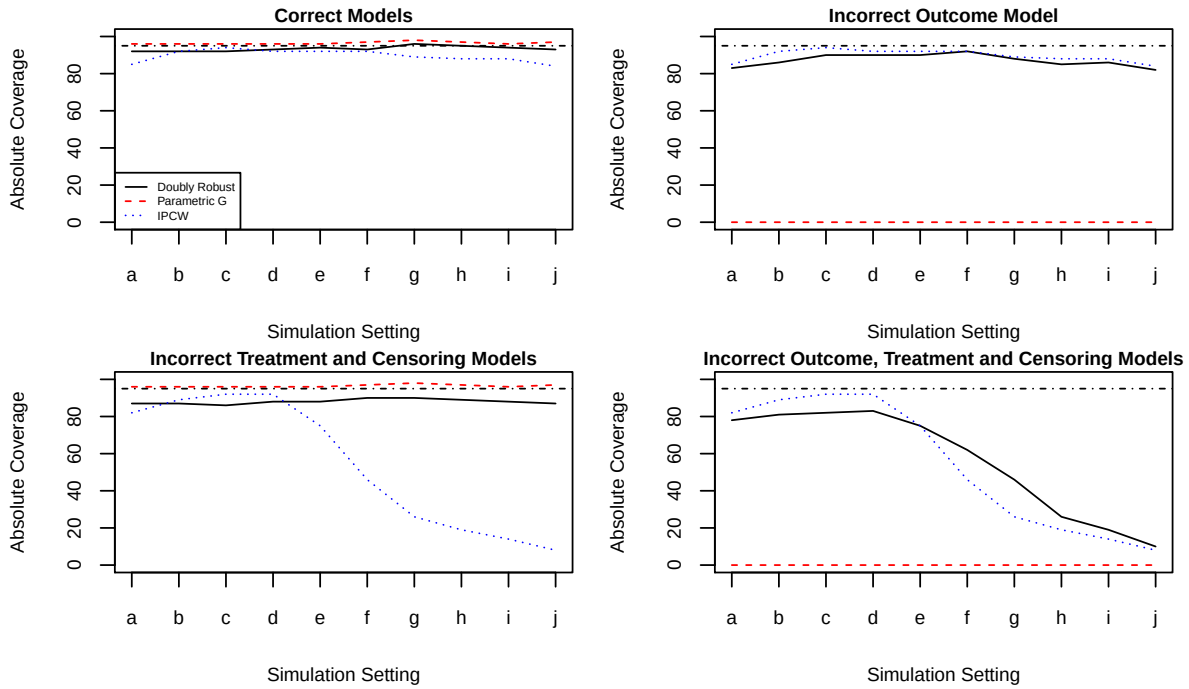


Figure 4.4: Coverages of the doubly robust, parametric g and the IPCW estimators under different model misspecifications in treatment group.

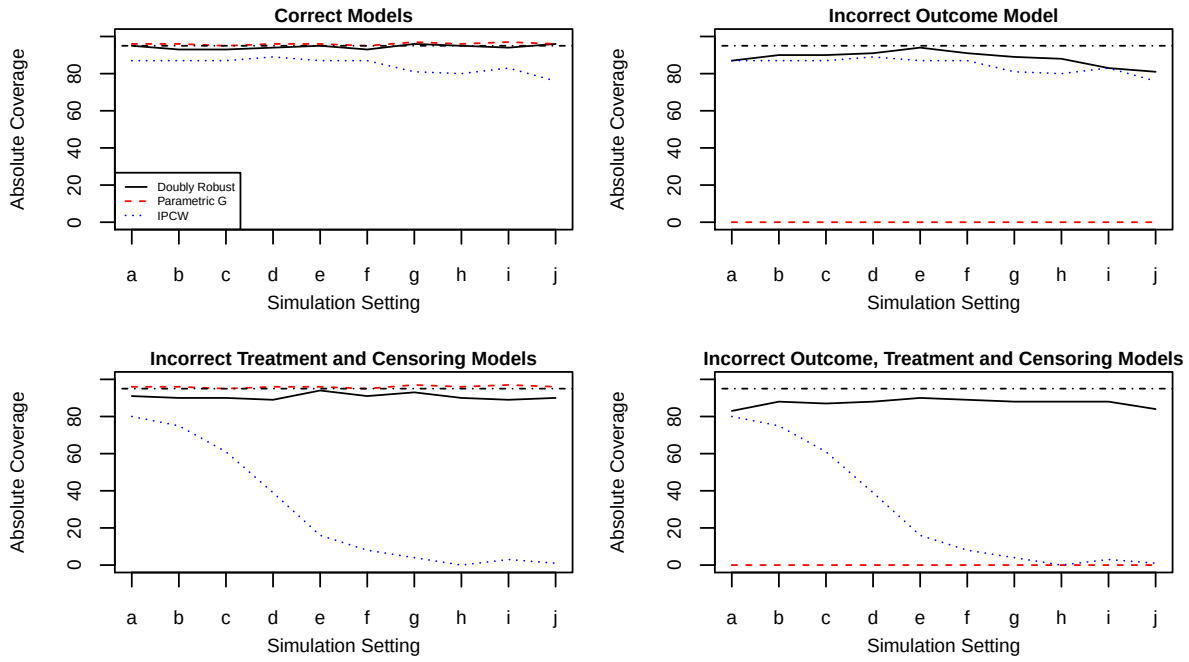
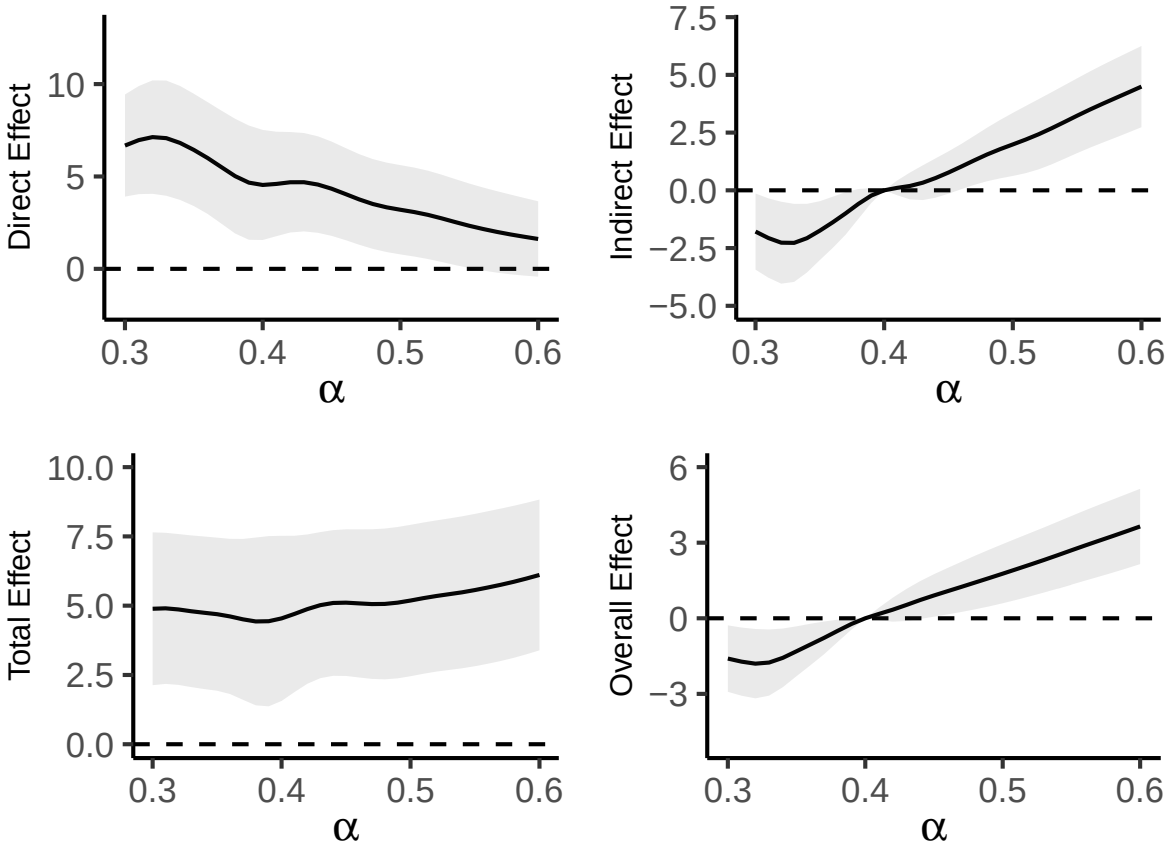


Figure 4.5: Direct effect, indirect effect, total effect and overall effect estimates multiplied by 1000 for different allocation strategies at time $t = 1$ year. Indirect effects, total effects and overall effects are with respect to $\alpha_2 = 0.4$. The shaded region denotes the 95% confidence interval of the estimates.



α	$\mu(100,0,\alpha)$	Bias	ESE	ASE	EC	α	$\mu(100,1,\alpha)$	Bias	ESE	ASE	EC
0.1	0.10	0.02	0.36	0.03	95%	0.1	0.43	0.00	0.13	0.02	97%
0.2	0.14	0.02	0.36	0.04	97%	0.2	0.48	0.00	0.18	0.03	95%
0.3	0.20	0.01	0.13	0.03	96%	0.3	0.53	0.00	0.09	0.03	95%
0.4	0.25	-0.01	0.03	0.02	94%	0.4	0.57	0.00	0.03	0.02	95%
0.5	0.31	-0.01	0.01	0.01	93%	0.5	0.61	0.00	0.01	0.01	95%
0.6	0.37	0.00	0.01	0.01	94%	0.6	0.65	0.00	0.01	0.01	94%
0.7	0.43	0.00	0.01	0.01	95%	0.7	0.68	0.00	0.01	0.01	94%
0.8	0.48	0.00	0.01	0.01	95%	0.8	0.71	0.00	0.01	0.01	95%
0.9	0.53	0.00	0.05	0.03	96%	0.9	0.74	0.00	0.03	0.01	96%

Table 4.1: Results from simulation study described in Section 4.3. α denotes the allocation probabilities, $\mu(100, a, \alpha)$ is the true value of the target parameter for $a = 0, 1$; Bias is the average of $\mu(100, a, \alpha) - \hat{F}^{DR}(100, a, \alpha)$ for $a = 0, 1$; ESE is the empirical standard error, ASE is the average of the sandwich variance estimates and EC denotes the empirical coverage of the 95% Wald confidence intervals.

	Gamma	Inverse Gaussian	Positive Stable
Exponential	5946.42 (6033.51)	5955.73 (6042.83)	5992.52 (6079.61)
Weibull	5888.98 (5985.76)	5898.34 (5995.12)	5935.18 (6031.96)
Gompertz	5858.04 (5954.82)	5867.50 (5964.28)	5904.57 (6001.35)
Loglogistic	5889.35 (5986.13)	5898.70 (5995.48)	5935.41 (6032.19)
Lognormal	7296.60 (7393.38)	7641.41 (7738.18)	6559.83 (6656.61)

Table 4.2: AIC (BIC) values for different baseline hazard functions corresponding to gamma, inverse Gaussian and positive stable frailty distributions.

CHAPTER 5: CONCLUSION

In the field of public health, interest often lies in estimating the effect of a treatment on an outcome of interest. In the causal inference framework, under the stable unit treatment value assumption (SUTVA), the causal effect of a treatment on the outcome can be used as a metric for the effect of treatment. One of the assumptions in SUTVA states that there is no interference. Interference is said to be present when the outcome of one individual is affected by the treatment status of another individual. A special case of interference is partial interference where it is assumed that interference can occur only between individuals within particular pre-specified groups but not between individuals of separate groups. In this document, we propose three different methods for estimating various effects of treatment on outcomes of interest in the presence of partial interference and right censoring. The different effects include direct effect of treatment as well as spillover effects due to partial interference.

Tchetgen and VanderWeele (2012) (TV) proposed IPW estimators for causal effects in the presence of partial interference. But they did not consider the nuances of censoring in their paper. We considered an extension of the TV IPW estimators by introducing a censoring weight along with the group propensity weights, previously used by TV. Following TV, the group propensity scores were obtained by assuming a mixed effects model for the treatment. The censoring weights were obtained by fitting a parametric frailty model to censoring times. Using the M-estimation theory, we proved that the estimator is consistent and asymptotically Normal. We performed simulation studies to show that the estimator had a very small bias for finite samples. Also, the sandwich variance estimator of the asymptotic standard error achieved the expected level of 95%

coverage.

The second method proposed for estimating causal effects was parametric g formula. The times to events were assumed to follow a parametric frailty model. Standardization was used to calculate the marginal survival probabilities. Consistency and asymptotic Normality of the estimator was proved as before using M-estimation theory. Again, simulation studies showed that the bias of the estimator was small and a sandwich variance estimator of the asymptotic standard error achieved nominal coverage.

Finally, we combined the IPW and parametric g formula estimators to propose a doubly robust estimator. The doubly robust estimator in the presence of partial interference and right censoring was obtained by extending the doubly robust estimator proposed by Liu et al. (2018) in the presence of partial interference only. We showed that the estimator was robust under model misspecification, i.e., the estimator was consistent and asymptotically Normal even if only one of the two sets of models were specified correctly, the outcome model or the censoring and treatment models. Simulation studies were again useful in showing the finite sample efficacy of the method as before.

All of the three methods discussed in the document were applied to a cholera vaccine study performed in Matlab, Bangladesh. Different methods yielded slightly different results but all of the methods suggested the presence of significant direct as well as spillover effects in the data. The methods also agreed upon the fact that the effect of treatment gets more pronounced over time. So, in accordance with previous studies, we conclude that vaccination has direct as well as indirect effect on the incidence of cholera in a particular neighborhood. Keeping everything else fixed, more vaccination should yield a lower number of cholera infections in a region.

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 2

Proof of proposition 1. From the definition of the IPCW estimator,

$$E\{\hat{F}_i(t,a,\alpha)\} = E\left\{\sum_{j=1}^{n_i} \frac{\pi(\mathbf{A}_{i,-j};\alpha)I(A_{ij}=a)I(\Delta_{ij}=1)I(X_{ij}\leq t)}{\Pr(\Delta_{ij}=1|\mathbf{L}_i,X_{ij})\Pr(\mathbf{A}_i|\mathbf{L}_i)n_i}\right\} \tag{1}$$

By the law of total expectation and causal consistency, the right side of (1) can be expressed as

$$E_{T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}E_{C_{ij}|T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}\left[\sum_{j=1}^{n_i} \frac{\pi(\mathbf{A}_{i,-j};\alpha)I(A_{ij}=a)I\{C_{ij}>T_{ij}(\mathbf{A}_i)\}I(T_{ij}(\mathbf{A}_i)\leq t)}{\Pr\{C_{ij}>T_{ij}(\mathbf{A}_i)|\mathbf{L}_i,T_{ij}(\mathbf{A}_i)\}\Pr(\mathbf{A}_i|\mathbf{L}_i)n_i}\right]$$

Moving the inner expectation inside the summation and taking out terms that are constant with respect to that expectation, it follows that

$$E\{\hat{F}_i(t,a,\alpha)\} = E_{T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}\left[\frac{\sum_{j=1}^{n_i}\pi(\mathbf{A}_{i,-j};\alpha)I(A_{ij}=a)I\{T_{ij}(\mathbf{A}_i)\leq t\}}{\Pr(\mathbf{A}_i|\mathbf{L}_i)n_i}\times \frac{E_{C_{ij}|T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}I\{C_{ij}>T_{ij}(\mathbf{A}_i)\}}{\Pr\{C_{ij}>T_{ij}(\mathbf{A}_i)|\mathbf{L}_i,T_{ij}(\mathbf{A}_i)\}}\right] \tag{2}$$

Next note that Assumption III implies that $\mathbf{C}_{ij} \perp\!\!\!\perp \{T_{ij}(\mathbf{A}_i),\mathbf{A}_i\}|\mathbf{L}_i$, which implies $\mathbf{C}_{ij} \perp\!\!\!\perp \mathbf{A}_i|\{T_{ij}(\mathbf{A}_i),\mathbf{L}_i\}$. Therefore,

$$\begin{aligned} E_{C_{ij}|T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}I\{C_{ij}>T_{ij}(\mathbf{A}_i)\} &= E_{C_{ij}|T_{ij}(\mathbf{A}_i),\mathbf{L}_i}I\{C_{ij}>T_{ij}(\mathbf{A}_i)\} \\ &= \Pr\{C_{ij}>T_{ij}(\mathbf{A}_i)|\mathbf{L}_i,T_{ij}(\mathbf{A}_i)\} \end{aligned}$$

implying (2) simplifies to

$$E\{\hat{F}_i(t,a,\alpha)\} = E_{T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}\left[\frac{\sum_{j=1}^{n_i}\pi(\mathbf{A}_{i,-j};\alpha)I(A_{ij}=a)I\{T_{ij}(\mathbf{A}_i)\leq t\}}{\Pr(\mathbf{A}_i|\mathbf{L}_i)n_i}\right] \tag{3}$$

Then, as in Tchetgen Tchetgen and VanderWeele (2012), it follows that

$$E\{\hat{F}_i(t,a,\alpha)\} = E_{\mathbf{T}_i(.),\mathbf{L}_i} E_{\mathbf{A}_i|\mathbf{T}_i(.),\mathbf{L}_i}\left[\frac{\sum_{j=1}^{n_i}\pi(\mathbf{A}_{i,-j};\alpha)I(A_{ij}=a)I\{T_{ij}(\mathbf{A}_i)\le t\}}{\Pr(\mathbf{A}_i|\mathbf{L}_i)n_i}\right]$$
$$= E_{\mathbf{T}_i(.),\mathbf{L}_i}\left[n_i^{-1}\sum_{j=1}^{n_i}\sum_{\mathbf{s}\in A(n_i-1)}\pi(\mathbf{s};\alpha)I\{T_{ij}(a_{ij}=a,\mathbf{a}_{i,-j}=\mathbf{s})\le t\}\times \left\{\frac{\Pr(A_{ij}=a,\mathbf{A}_{i,-j}=\mathbf{s}|\mathbf{T}_i(.),\mathbf{L}_i)}{\Pr(A_{ij}=a,\mathbf{A}_{i,-j}=\mathbf{s}|\mathbf{L}_i)}\right\}\right]$$

By assumption I, $\Pr(A_{ij}=a,\mathbf{A}_{i,-j}=\mathbf{s}|\mathbf{T}_i(.),\mathbf{L}_i) = \Pr(A_{ij}=a,\mathbf{A}_{i,-j}=\mathbf{s}|\mathbf{L}_i)$. Therefore

$$E\{\hat{F}_i(t,a,\alpha)\} = E\left[n_i^{-1}\sum_{j=1}^{n_i}\sum_{\mathbf{s}\in A(n_i-1)}\pi(\mathbf{s};\alpha)I\{T_{ij}(a_{ij}=a,\mathbf{a}_{i,-j}=\mathbf{s})\le t\}\right]$$
$$= E\{\bar{F}_i(t,a,\alpha)\} = \mu(t,a,\alpha).$$

A similar proof can be used to show $E\{\hat{\mu}(t,\alpha)\} = \mu(t,\alpha)$.

APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 3

Proof of proposition 2. According to the definition of the parametric g estimator

$$m_i^{int}(t,a,\alpha,\boldsymbol{\omega}) = \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_{i,-j}\in\mathcal{A}(n_i-1)} \Pr(T_{ij}\le t|\mathbf{L}_{ij}, A=a, \mathbf{A}_{i,-j}=\mathbf{a}_{i,-k},\boldsymbol{\omega})\pi(\mathbf{a}_{i,-j},\alpha)$$

Using causal consistency,

$$\Pr(T_{ij}\le t|\mathbf{L}_i, A=a, \mathbf{A}_{i,-j}=\mathbf{a}_{i,-j},\boldsymbol{\omega})$$
$$= \Pr(T_{ij}(a,\mathbf{a}_{i,-j})\le t|\mathbf{L}_i, A=a, \mathbf{A}_{i,-j}=\mathbf{a}_{i,-j},\boldsymbol{\omega})$$

Then, using the conditional independence assumption,

$$\Pr(T_{ij}(a,\mathbf{a}_{i,-k})\le t|\mathbf{L}_i, A=a, \mathbf{A}_{i,-j}=\mathbf{a}_{i,-j},\boldsymbol{\omega}) = \Pr(T_{ij}(a,\mathbf{a}_{i,-j})\le t|\mathbf{L}_i,\boldsymbol{\omega})$$

$$= E_{\mathbf{T}_i(.)|\mathbf{L}_i} I(T_{ij}(a,\mathbf{a}_{i,-k})\le t)$$

So,

$$E_{\mathbf{L}_i}\{m_i^{int}(t,a,\alpha,\boldsymbol{\omega})\} = E_{\mathbf{L}_i}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_{i,-j}\in\mathcal{A}(n_i-1)} E_{\mathbf{T}_i(.)|\mathbf{L}_i} I(T_{ij}(a,\mathbf{a}_{i,-k})\le t)\pi(\mathbf{a}_{i,-j},\alpha)$$

$$= E_{\mathbf{L}_i}E_{\mathbf{T}_i(.)|\mathbf{L}_i}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_{i,-j}\in\mathcal{A}(n_i-1)} I(T_{ij}(a,\mathbf{a}_{i,-k})\le t)\pi(\mathbf{a}_{i,-j},\alpha)$$

$$= E_{\mathbf{T}_i(.),\mathbf{L}_i}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{\mathbf{a}_{i,-j}\in\mathcal{A}(n_i-1)} I(T_{ij}(a,\mathbf{a}_{i,-k})\le t)\pi(\mathbf{a}_{i,-j},\alpha)$$

$$= \mu(t,a,\alpha)$$

Hence,

$$E\{m_i^{int}(t,a,\alpha,\boldsymbol{\omega})\} = \mu(t,a,\alpha)$$

Similar calculations will yield

$$E\{m_i^{int}(t, \alpha, \boldsymbol{\omega})\} = \mu(t, \alpha)$$

APPENDIX C: TECHNICAL DETAILS FOR CHAPTER 4

Proof of proposition 3. Assume $\boldsymbol{\beta}^*$, $\boldsymbol{\gamma}^*$, and $\boldsymbol{\omega}^*$ to be such that $E\{\psi_{ck}(\mathbf{X}_i, \boldsymbol{\Delta}_i, \mathbf{L}_i, \boldsymbol{\gamma}^*)\} = E\{\psi_{xk}(\mathbf{A}_i, \mathbf{L}_i, \boldsymbol{\beta}^*)\} = E\{\psi_{ck}^{OR}(\mathbf{X_i}, \boldsymbol{\Delta_i}, \mathbf{L_i}, \boldsymbol{\omega}^*)\} = 0$ where the expectation is taken over the true parameter values. The following proof shows that $E(\psi_{a\alpha}^{DR}(\mathbf{O}_i, \boldsymbol{\theta}^{DR*})) = 0$ when either $\boldsymbol{\omega}^* = \boldsymbol{\omega}_0$ or $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ where $\boldsymbol{\theta}^{DR*} = (\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*, \boldsymbol{\omega}^*, \mu(t, a, \alpha))$. Now,

$$E\Bigg(n_i^{-1}\sum_{j=1}^{n_i}\Bigg[\frac{I(A_{ij}=a)I(\Delta_{ij}^t=1)\{I(X_{ij}\le t)-m_{ij}(\mathbf{A}_i,t,\mathbf{L}_i,\boldsymbol{\omega}^*)\}\pi(\mathbf{A}_{i,-j};\alpha)}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\boldsymbol{\beta}^*)\Pr(\Delta_{ij}^t=1|\mathbf{L}_i,\boldsymbol{\gamma}^*,X_{ij})} + \sum_{a_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\pi(\mathbf{a}_{i,-j};\alpha)\Bigg]\Bigg)$$

Using the law of total expectation and taking out terms that are constant with respect to the expectation $E_{C_{ij}|T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}$,

$$= E_{T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}\Bigg(n_i^{-1}\sum_{j=1}^{n_i}\Big[\frac{I(A_{ij}=a)\{I(T_{ij}(\mathbf{A}_i)\le t)-m_{ij}(\mathbf{A}_i,t,\mathbf{L}_i,\boldsymbol{\omega}^*)\}\pi(\mathbf{A}_{i,-j};\alpha)}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\boldsymbol{\beta}^*)} \times \frac{E_{C_{ij}|T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}I\{C_{ij}>T_{ij}^t(\mathbf{A}_i)\}}{\Pr\{C_{ij}>T_{ij}^t(\mathbf{A}_i)|\mathbf{L}_i,T_{ij}(\mathbf{A}_i),\boldsymbol{\gamma}^*\}}\Big]\Bigg) + E\left\{n_i^{-1}\sum_{j=1}^{n_i}\sum_{a_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\pi(\mathbf{a}_{i,-j};\alpha)\right\} \quad (4)$$

Correct treatment and censoring models

Note that Assumption III implies that $C_{ij} \perp\!\!\!\perp \{T_{ij}(\mathbf{A}_i), \mathbf{A}_i\}|\mathbf{L}_i$, which implies $C_{ij} \perp\!\!\!\perp \mathbf{A}_i|\{T_{ij}(\mathbf{A}_i), \mathbf{L}_i\}$. If $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$ then

$$\frac{E_{C_{ij}|T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}I\{C_{ij}>T_{ij}^t(\mathbf{A}_i)\}}{\Pr\{C_{ij}>T_{ij}^t(\mathbf{A}_i)|\mathbf{L}_i,T_{ij}(\mathbf{A}_i),\boldsymbol{\gamma}^*\}} = 1$$

because

$$E_{C_{ij}|T_{ij}^t(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i} I\{C_{ij} > T_{ij}^t(\mathbf{A}_i)\} = E_{C_{ij}|T_{ij}(\mathbf{A}_i),\mathbf{L}_i} I\{C_{ij} > T_{ij}(\mathbf{A}_i)\}$$
$$= \Pr\{C_{ij} > T_{ij}^t(\mathbf{A}_i)|\mathbf{L}_i, T_{ij}(\mathbf{A}_i), \gamma_0\}$$

So, the term in (4) equals

$$= E_{T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}\left(n_i^{-1}\sum_{j=1}^{n_i}\Big[\frac{I(A_{ij}=a)\{I(T_{ij}(\mathbf{A}_i)\le t) - m_{ij}(\mathbf{A}_i,t,\mathbf{L}_i,\boldsymbol{\omega}^*)\}}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\boldsymbol{\beta}^*)} \times \pi(\mathbf{A}_{i,-j};\alpha)\Big]\right) + E\left\{ n_i^{-1}\sum_{j=1}^{n_i}\sum_{a_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\pi(\mathbf{a}_{i,-j};\alpha)\right\}$$

Again using the law of total expectation and exchanging sums,

$$= E_{T_{ij}(\mathbf{A}_i),\mathbf{L}_i}\left(n_i^{-1}\sum_{j=1}^{n_i} E_{\mathbf{A}_i|T_{ij}(\mathbf{A}_i),\mathbf{L}_i}\Big[\frac{I(A_{ij}=a)\{I(T_{ij}(\mathbf{A}_i)\le t) - m_{ij}(\mathbf{A}_i,t,\mathbf{L}_i,\boldsymbol{\omega}^*)\}}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\boldsymbol{\beta}^*)} \times \pi(\mathbf{A}_{i,-j};\alpha)\Big]\right) + E\left\{ n_i^{-1}\sum_{j=1}^{n_i}\sum_{a_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\pi(\mathbf{a}_{i,-j};\alpha)\right\}$$

Replacing the expectation with sum,

$$= E_{\mathbf{T}_i(.),\mathbf{L}_i}\left(n_i^{-1}\sum_{j=1}^{n_i}\sum_{\mathbf{s}\in A(n_i-1)} \pi(\mathbf{s};\alpha)\big[I\{T_{ij}(a_{ij}=a,\mathbf{a}_{i,-j}=\mathbf{s})\le t\} - m_{ij}(a,\mathbf{s},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\big]\left\{\frac{\Pr(A_{ij}=a,\mathbf{A}_{i,-j}=\mathbf{s}|\mathbf{T}_i(.),\mathbf{L}_i,\boldsymbol{\beta}_0)}{\Pr(A_{ij}=a,\mathbf{A}_{i,-j}=\mathbf{s}|\mathbf{L}_i,\boldsymbol{\beta}^*)}\right\}\right) + E\left\{ n_i^{-1}\sum_{j=1}^{n_i}\sum_{a_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\pi(\mathbf{a}_{i,-j};\alpha)\right\} \quad (5)$$

If $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, then using conditional independence, the term in (5) equals,

$$= E_{\mathbf{T}_i(.),\mathbf{L}_i}\Bigg(n_i^{-1}\sum_{j=1}^{n_i}\sum_{\mathbf{s}\in A(n_i-1)} \pi(\mathbf{s};\alpha)\big[I\left\{T_{ij}(a_{ij}=a,\mathbf{a}_{i,-j}=\mathbf{s})\le t\right\} - m_{ij}(a,\mathbf{s},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\big]\Bigg) + E\left\{ n_i^{-1}\sum_{j=1}^{n_i}\sum_{a_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\pi(\mathbf{a}_{i,-j};\alpha)\right\}$$

$$= E_{\mathbf{T}_i(.),\mathbf{L}_i}\left(n_i^{-1}\sum_{j=1}^{n_i}\sum_{\mathbf{s}\in A(n_i-1)} \pi(\mathbf{s};\alpha) I\left\{T_{ij}(a_{ij}=a,\mathbf{a}_{i,-j}=\mathbf{s})\le t\right\}\right)$$

$$= \mu(t,a,\alpha)$$

So, $E(\psi^{DR}_{a\alpha}(\mathbf{O}_i,\boldsymbol{\theta}^{DR*}))=0$ when $\boldsymbol{\beta}^*=\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}^*=\boldsymbol{\gamma}_0$.

Correct outcome model

As shown before, the term in (4.1) can be written as,

$$= E_{T_{ij}(\mathbf{A}_i),\mathbf{A}_i,\mathbf{L}_i}\Bigg(n_i^{-1}\sum_{j=1}^{n_i}\Big[\frac{I(A_{ij}=a)\{I(T_{ij}(\mathbf{A}_i)\le t)-m_{ij}(\mathbf{A}_i,t,\mathbf{L}_i,\boldsymbol{\omega}^*)\}\pi(\mathbf{A}_{i,-j};\alpha)}{\Pr(\mathbf{A}_i|\mathbf{L}_i,\boldsymbol{\beta}^*)} \times \frac{\Pr\{C_{ij}>T^t_{ij}(\mathbf{A}_i)|\mathbf{L}_i,\boldsymbol{\gamma}_0\}}{\Pr\{C_{ij}>T^t_{ij}(\mathbf{A}_i)|\mathbf{L}_i,\boldsymbol{\gamma}^*\}}\Big]\Bigg) + E\left\{ n_i^{-1}\sum_{j=1}^{n_i}\sum_{a_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\pi(\mathbf{a}_{i,-j};\alpha)\right\}$$

Again using law of total expectation as before and replacing the expectation with sum,

$$= E_{\mathbf{L}_i}\Bigg(n_i^{-1}\sum_{j=1}^{n_i}\sum_{\mathbf{s}\in A(n_i-1)} \pi(\mathbf{s};\alpha)\big[E_{\mathbf{T}_i(.)|\mathbf{L}_i}\left\{I\left\{T_{ij}(a_{ij}=a,\mathbf{a}_{i,-j}=\mathbf{s})\le t\right\}\right\} - m_{ij}(a,\mathbf{s},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\big] \times \left\{\frac{\Pr(A_{ij}=a,\mathbf{A}_{i,-j}=\mathbf{s}|\mathbf{L}_i,\boldsymbol{\beta}_0)}{\Pr(A_{ij}=a,\mathbf{A}_{i,-j}=\mathbf{s}|\mathbf{L}_i,\boldsymbol{\beta}^*)}\right\} \times \frac{\Pr\{C_{ij}>T^t_{ij}(\mathbf{A}_i)|\mathbf{L}_i,\boldsymbol{\gamma}_0\}}{\Pr\{C_{ij}>T^t_{ij}(\mathbf{A}_i)|\mathbf{L}_i,\boldsymbol{\gamma}^*\}}\Bigg) + E\left\{ n_i^{-1}\sum_{j=1}^{n_i}\sum_{a_{i,-j}} m_{ij}(a,\mathbf{a}_{i,-j},t,\mathbf{L}_i,\boldsymbol{\omega}^*)\pi(\mathbf{a}_{i,-j};\alpha)\right\} \quad (6)$$

If $\boldsymbol{\omega}^* = \boldsymbol{\omega}_0$ then $m_{ij}(a,\mathbf{s},t,\mathbf{L}_i,\boldsymbol{\omega}^*) = E_{\mathbf{T}_i(.)|\mathbf{L}_i}\left\{I\left\{T_{ij}(a_{ij}=a,\mathbf{a}_{i,-j}=\mathbf{s})\le t\right\}\right\}$. So, the

term in (6) equals

$$E\left\{ n_i^{-1} \sum_{j=1}^{n_i} \sum_{a_{i,-j}} m_{ij}(a, \mathbf{a}_{i,-j}, t, \mathbf{L}_i, \boldsymbol{\omega}_0) \pi(\mathbf{a}_{i,-j}; \alpha) \right\} = \mu(t, a, \alpha)$$

So, $E(\psi_{a\alpha}^{DR}(\mathbf{O}_i, \boldsymbol{\theta}^{DR*})) = 0$ when $\boldsymbol{\omega}^* = \boldsymbol{\omega}_0$ or when $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$. By M-estimation theory (Stefanski and Boos 2002), $\sqrt{m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{DR*})$ converges in distribution to a normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}^{DR}$ equal to $U(\boldsymbol{\theta}^{DR*})^{-1} V(\boldsymbol{\theta}^{DR*}) \{U(\boldsymbol{\theta}^{DR*})^{-1}\}^T$ where $U(\boldsymbol{\theta}^{DR*}) = E\{-\dot{\psi}(\mathbf{O}_i, \boldsymbol{\theta})\}$, $V(\boldsymbol{\theta}^{DR*}) = E\{\psi(\mathbf{O}_i, \boldsymbol{\theta}^{DR*}) \psi(\mathbf{O}_i, \boldsymbol{\theta}^{DR*})^T\}$, and $\dot{\psi}(\mathbf{O}_i, \boldsymbol{\theta}) = \partial \psi(\mathbf{O}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T$ when $\boldsymbol{\omega}^* = \boldsymbol{\omega}_0$ or when $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}_0$. Consistency and asymptotic normality of the direct, indirect and total effect estimators follows from the delta method. Similar techniques can be used to show that $\hat{\mu}(t, \alpha)$ and the overall effect estimator are also consistent and asymptotically normal.

BIBLIOGRAPHY

Ali, M., Emch, M., von Seidlein, L., Yunus, M., Sack, D. A., Rao, M., Holmgren, J., and Clemens, J. D. (2005), "Herd immunity conferred by killed oral cholera vaccines in Bangladesh: a reanalysis," *The Lancet*, 366, 44–49.

Aronow, P. M. and Samii, C. (2017), "Estimating average causal effects under interference between units," *Annals of Applied Statistics*, 11, 1912–1947.

Baird, S., Bohren, J., McIntosh, C., and Özler, B. (2018), "Optimal design of experiments in the presence of interference," *Review of Economics and Statistics*, in press.

Bang, H. and Robins, J. M. (2005), "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–973.

Berg, G. J. and Drepper, B. (2012), "Inference for shared-frailty survival models with left-truncated data," Tech. rep., Working Paper Series, Department of Economics, University of Mannheim.

Bowers, J., Fredrickson, M. M., and Panagopoulos, C. (2013), "Reasoning about interference between units: A general framework," *Political Analysis*, 21, 97–124.

Cain, L. E. and Cole, S. R. (2009), "Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death," *Statistics in Medicine*, 28, 1725–1738.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009), "Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data," *Biometrika*, 96, 723–734.

Cassel, C., Särndal, C.-E., and Wretman, J. (1977), *Foundations of Inference in Survey Sampling*, Wiley New York.

Clayton, D. G. (1991), "A Monte Carlo method for Bayesian inference in frailty models," *Biometrics*, 467–485.

Cole, S. R. and Frangakis, C. E. (2009), "The consistency statement in causal inference: a definition or an assumption?" *Epidemiology*, 20, 3–5.

Cole, S. R. and Hernán, M. A. (2008), "Constructing inverse probability weights for marginal structural models," *American Journal of Epidemiology*, 168, 656–664.

Cole, S. R., Richardson, D. B., Chu, H., and Naimi, A. I. (2013), "Analysis of occupational asbestos exposure and lung cancer mortality using the g formula," *American Journal of Epidemiology*, 177, 989–996.

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.

Duncan, G. J., Boisjoly, J., Kremer, M., Levy, D. M., and Eccles, J. (2005), "Peer effects in drug use and sex among college students," *Journal of Abnormal Child Psychology*, 33, 375–385.

Eckles, D., Karrer, B., and Ugander, J. (2014), "Design and analysis of experiments in networks: Reducing bias from interference," *arXiv preprint arXiv:1404.7530*.

Eckles, D., Kizilcec, R. F., and Bakshy, E. (2016), "Estimating peer effects in networks with peer encouragement designs," *Proceedings of the National Academy of Sciences*, 113, 7316–7322.

Emch, M., Ali, M., Park, J.-K., Yunus, M., Sack, D. A., and Clemens, J. D. (2006), "Relationship between neighbourhood-level killed oral cholera vaccine coverage and protective efficacy: evidence for herd immunity," *International Journal of Epidemiology*, 35, 1044–1050.

Foster, E. M. (2010), "Causal inference and developmental psychology." *Developmental Psychology*, 46, 1454.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011), "Doubly robust estimation of causal effects," *American Journal of Epidemiology*, 173, 761–767.

Garcia-Aymerich, J., Varraso, R., Danaei, G., Camargo Jr, C. A., and Hernán, M. A. (2013), "Incidence of adult-onset asthma after hypothetical interventions on body mass index and physical activity: an application of the parametric g-formula," *American Journal of Epidemiology*, 179, 20–26.

Gauderman, W. J. and Thomas, D. C. (1994), "Censored survival models for genetic epidemiology: a Gibbs sampling approach," *Genetic Epidemiology*, 11, 171–188.

Gibbard, A. and Harper, W. L. (1976), "Counterfactuals and Two Kinds of," .

Glidden, D. V. and Vittinghoff, E. (2004), "Modelling clustered survival data from multicentre clinical trials," *Statistics in Medicine*, 23, 369–388.

Graham, D. J., McCoy, E. J., and Stephens, D. A. (2013), "Quantifying the effect of area deprivation on child pedestrian casualties by using longitudinal mixed models to adjust for confounding, interference and spatial dependence," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176, 931–950.

Greenland, S. and Robins, J. M. (1986), "Identifiability, exchangeability, and epidemiological confounding," *International Journal of Epidemiology*, 15, 413–419.

Gutierrez, R. G. et al. (2002), "Parametric frailty and shared frailty survival models," *Stata Journal*, 2, 22–44.

Haavelmo, T. (1944), "The probability approach in econometrics," *Econometrica: Journal of the Econometric Society*, iii–115.

Halloran, M. E. and Struchiner, C. J. (1991), "Study designs for dependent happenings," *Epidemiology*, 331–338.

— (1995), "Causal inference in infectious diseases," *Epidemiology*, 142–151.

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000), "Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men," *Epidemiology*, 11, 561–570.

Hernán, M. A. and Robins, J. M. (2006), "Estimating causal effects from epidemiological data," *Journal of Epidemiology & Community Health*, 60, 578–586.

Hernan, M. A. and Robins, J. M. (2010), *Causal Inference*, CRC Boca Raton, FL:.

Holt, J. and Prentice, R. (1974), "Survival analyses in twin studies and matched pair experiments," *Biometrika*, 61, 17–30.

Hong, G. and Raudenbush, S. W. (2006), "Evaluating kindergarten retention policy, A Case Study of Causal Inference for Multilevel Observational Data," *Journal of the American Statistical Association*, 101, 901–910.

Horvitz, D. G. and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American statistical Association*, 47, 663–685.

Hougaard, P. (1984), "Life table methods for heterogeneous populations: Distributions describing the heterogeneity," *Biometrika*, 71, 75–83.

Howe, C. J., Cole, S. R., Chmiel, J. S., and Muñoz, A. (2011), "Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias," *American Journal of Epidemiology*, 173, 569–577.

Hudgens, M. G. and Halloran, M. E. (2008), "Toward causal inference with interference," *Journal of the American Statistical Association*, 103, 832–842.

Kang, J. D. and Schafer, J. L. (2007), "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical Science*, 523–539.

Keil, A. P., Edwards, J. K., Richardson, D. R., Naimi, A. I., and Cole, S. R. (2014), "The parametric G-formula for time-to-event data: towards intuition with a worked example," *Epidemiology (Cambridge, Mass.)*, 25, 889.

Korsgaard, I. R., Madsen, P., and Jensen, J. (1998), "Bayesian inference in the semiparametric log normal frailty model using Gibbs sampling," *Genetics Selection Evolution*, 30, 241.

Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014), "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, 111, 8788–8790.

Lancaster, T. (1979), "Econometric methods for the duration of unemployment," *Econometrica: Journal of the Econometric Society*, 939–956.

Lefebvre, G., Delaney, J. A., and Platt, R. W. (2008), "Impact of mis-specification of the treatment model on estimates from a marginal structural model," *Statistics in Medicine*, 27, 3629–3642.

Li, H. and Thompson, E. (1997), "Semiparametric estimation of major gene and family-specific random effects for age of onset," *Biometrics*, 282–293.

Li, H. and Zhong, X. (2002), "Multivariate survival models induced by genetic frailties, with application to linkage analysis," *Biostatistics*, 3, 57–75.

Little, R. J. and Rubin, D. B. (2014), *Statistical Analysis with Missing Data*, John Wiley & Sons.

Liu, L., Hudgens, M., and Becker-Dreps, S. (2016), "On inverse probability-weighted estimators in the presence of interference," *Biometrika*, 103, 829–842.

Liu, L., Hudgens, M. G., Saul, B., Clemens, J. D., Ali, M., and Emch, M. E. (2018), "Doubly Robust Estimation in Observational Studies with Partial Interference," *arXiv preprint arXiv:1806.07422*.

Longini, I. M., Yunus, M., Zaman, K., Siddique, A., Sack, R. B., and Nizam, A. (2002), "Epidemic and endemic cholera trends over a 33-year period in Bangladesh," *Journal of Infectious Diseases*, 186, 246–251.

Lunceford, J. K. and Davidian, M. (2004), "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in Medicine*, 23, 2937–2960.

Luo, X., Small, D. S., Li, C.-S. R., and Rosenbaum, P. R. (2012), "Inference with interference between units in an fMRI experiment of motor inhibition," *Journal of the American Statistical Association*, 107, 530–541.

Manski, C. F. (2013), "Identification of treatment response with social interactions," *The Econometrics Journal*, 16, S1–S23.

Munda, M., Rotolo, F., and Legrand, C. (2012), "parfm: Parametric frailty models in R," *Journal of Statistical Software*, 51, 1–20.

Naimi, A. I. and Kennedy, E. H. (2017), "Nonparametric double robustness," *arXiv preprint arXiv:1711.07137*.

Pankratz, V. S., De Andrade, M., and Therneau, T. M. (2005), "Random-effects Cox proportional hazards model: general variance components methods for time-to-event data," *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 28, 97–109.

Perez-Heydrich, C., Hudgens, M. G., Halloran, M. E., Clemens, J. D., Ali, M., and Emch, M. E. (2014), "Assessing effects of cholera vaccination in the presence of interference," *Biometrics*, 70, 731–741.

Pratt, J. W. and Schlaifer, R. (1984), "On the nature and discovery of structure," *Journal of the American Statistical Association*, 79, 9–21.

— (1988), "On the interpretation and observation of laws," *Journal of Econometrics*, 39, 23–52.

Robins, J. (1986), "A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393–1512.

— (1987), "A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods," *Journal of chronic diseases*, 40, 139S–161S.

Robins, J. M. (2000), "Marginal structural models versus structural nested models as tools for causal inference," in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, Springer, pp. 95–133.

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992), "G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients," *Epidemiology*, 319–336.

Robins, J. M. and Finkelstein, D. M. (2000), "Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests," *Biometrics*, 56, 779–788.

Robins, J. M. and Greenland, S. (1996), "Identification of causal effects using instrumental variables: comment," *Journal of the American Statistical Association*, 91, 456–458.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000), "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 11, 550–560.

Rosenbaum, P. R. (2007), "Interference between units in randomized experiments," *Journal of the American Statistical Association*, 102, 191–200.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rubin, D. B. (1980), "Randomization analysis of experimental data: The Fisher randomization test comment," *Journal of the American Statistical Association*, 75, 591–593.

— (2004), "Direct and indirect causal effects via potential outcomes," *Scandinavian Journal of Statistics*, 31, 161–170.

Sampson, R. J. (2010), "Gold standard myths: Observations on the experimental turn in quantitative criminology," *Journal of Quantitative Criminology*, 26, 489–500.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003), *Model assisted survey sampling*, Springer Science & Business Media.

Saul, B. C. and Hudgens, M. G. (2017), "The Calculus of M-estimation in R with geex," *arXiv preprint arXiv:1709.01413*.

Sinclair, B., McConnell, M., and Green, D. P. (2012), "Detecting spillover effects: Design and analysis of multilevel experiments," *American Journal of Political Science*, 56, 1055–1069.

Sobel, M. E. (1990), "Effect analysis and causation in linear structural equation models," *Psychometrika*, 55, 495–515.

— (1995), "Causal inference in the social and behavioral sciences," in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Springer, pp. 1–38.

— (2006), "What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference," *Journal of the American Statistical Association*, 101, 1398–1407.

Splawa-Neyman, J., Dabrowska, D., Speed, T., et al. (1923), "On the application of probability theory to agricultural experiments. Essay on principles. Section 9," *Statistical Science*, 5, 465–472.

Stefanski, L. A. and Boos, D. D. (2002), "The calculus of M-estimation," *The American Statistician*, 56, 29–38.

Taubman, S. L., Robins, J. M., Mittleman, M. A., and Hernán, M. A. (2009), "Intervening on risk factors for coronary heart disease: an application of the parametric g-formula," *International Journal of Epidemiology*, 38, 1599–1611.

Tchetgen, E. J. T. and VanderWeele, T. J. (2012), "On causal inference in the presence of interference," *Statistical Methods in Medical Research*, 21, 55–75.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003), "Penalized survival models and frailty," *Journal of Computational and Graphical Statistics*, 12, 156–175.

Toulis, P. and Kao, E. (2013), "Estimation of causal peer influence effects," *Proceedings of The 30th International Conference on Machine Learning*, 1489–1497.

VanderWeele, T. J. and An, W. (2013), "Social networks and causal inference," Springer, pp. 353–374.

Vanderweele, T. J., Hong, G., Jones, S. M., and Brown, J. L. (2013), "Mediation and spillover effects in group-randomized trials: a case study of the 4Rs educational intervention," *Journal of the American Statistical Association*, 108, 469–482.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979), "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography*, 16, 439–454.

Verbitsky-Savitz, N. and Raudenbush, S. W. (2012), "Causal inference under interference in spatial settings: A case study evaluating community policing program in Chicago," *Epidemiologic Methods*, 1, 107–130.

Wei, L.-J., Lin, D. Y., and Weissfeld, L. (1989), "Regression analysis of multivariate incomplete failure time data by modeling marginal distributions," *Journal of the American Statistical Association*, 84, 1065–1073.

Westreich, D. and Cole, S. R. (2010), "Invited commentary: positivity in practice," *American Journal of Epidemiology*, 171, 674–677.

Westreich, D., Cole, S. R., Young, J. G., Palella, F., Tien, P. C., Kingsley, L., Gange, S. J., and Hernán, M. A. (2012), "The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death," *Statistics in Medicine*, 31, 2000–2009.

Willkinson, L. (1999), "Statistical Methods in Psychology Journals." *American Psychologist*, 54, 594–604.

Young, J. G., Cain, L. E., Robins, J. M., O'Reilly, E. J., and Hernán, M. A. (2011), "Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula," *Statistics in Biosciences*, 3, 119.

Zigler, C. M., Dominici, F., and Wang, Y. (2012), "Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes," *Biostatistics*, 13, 289–302.